NAVAL
POSTGRADUATE
SCHOOL

MONTEREY, CALIFORNIA

DISSERTATION

MODELING HUMAN VISUAL PERCEPTION FOR TARGET DETECTION IN MILITARY SIMULATIONS

by

Patrick Jungkunz

June 2009

Dissertation Supervisor: Christian J. Darken

This dissertation was done at the MOVES Institute.
Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 2009	Dissertation

4. TITLE AND SUBTITLE: Modeling Human Visual Perception for Target Detection in Military Simulations

5. FUNDING NUMBERS

6. AUTHOR(S): Patrick W. Jungkunz

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES: The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)
The search and target acquisition models used in current military simulations for visual detection of ground soldiers are empirical. Although taking into account human performance data collected in field trials, they do not attempt to realistically model human search behavior. This, however, is necessary to achieve realistic target detection performance, including such phenomena as false positive detections at realistic locations. Working towards this goal, this research creates a model of human visual perception for the search of a human target. The contributions of bottom-up and top-down information on human visual perception are examined in a visual search experiment, which includes eye movement recording of the participants. The results show that semantically relevant scene information is used to guide the search process, influencing eye movements. Consequently, a predictive model of eye fixations is created which takes semantically relevant scene locations into account. These meaningful locations are extracted from ground truth simulation data and fused into a relevancy map. The relevance map is compared with eye fixations of participants searching for human targets in realistic scenes. This comparison shows that the relevance map predicts fixation locations very well. A combination of the relevance map with a salience map achieves even better prediction of eye fixations.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Human Visual Perception, Visual Attention, Eye Movements, Eye Tracking, Human Behavior Modeling, Target Detection, Visual Search, Semantic Relevance, Relevance Map	172
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

MODELING HUMAN VISUAL PERCEPTION FOR TARGET DETECTION IN MILITARY SIMULATIONS

Patrick W. Jungkunz
Kapitänleutnant, German Navy
Dipl.-Inform. (univ.), Universtität der Bundeswehr München, 2000

Submitted in partial fulfillment of the
requirements for the degree of

**DOCTOR OF PHILOSOPHY IN
MODELING, VIRTUAL ENVIRONMENTS AND SIMULATION**

from the

**NAVAL POSTGRADUATE SCHOOL
June 2009**

Author: Patrick W. Jungkunz

Approved By:

Christian J. Darken Associate Professor of Computer Science Dissertation Supervisor	Rudolph P. Darken Professor of Computer Science
Anthony Ciavarelli Research Professor of MOVES	Thomas W. Lucas Associate Professor of Operations Research
John E. Hiles Research Professor of Computer Science	Kevin Squire Assistant Professor of Computer Science

Approved By: Mathias Kölsch, Chair, MOVES Academic Committee

Approved By: Doug Moses, Vice Provost for Academic Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The search and target acquisition models used in current military simulations for visual detection of ground soldiers are empirical. Although taking into account human performance data collected in field trials, they do not attempt to realistically model human search behavior. This, however, is necessary to achieve realistic target detection performance, including such phenomena as false positive detections at realistic locations. Working towards this goal, this research creates a model of human visual perception for the search of a human target. The contributions of bottom-up and top-down information on human visual perception are examined in a visual search experiment, which includes eye movement recording of the participants. The results show that semantically relevant scene information is used to guide the search process, influencing eye movements. Consequently, a predictive model of eye fixations is created which takes semantically relevant scene locations into account. These meaningful locations are extracted from ground truth simulation data and fused into a relevancy map. The relevance map is compared with eye fixations of participants searching for human targets in realistic scenes. This comparison shows that the relevance map predicts fixation locations very well. A combination of the relevance map with a salience map achieves even better prediction of eye fixations.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGEMENTS

Without the numerous people assisting this dissertation, it would not be what it is. First of all, I am deeply grateful to my advisor, Chris Darken, who supported me in any way he could all the way from finding a topic up to finishing this dissertation. His guidance and his advice were essential, his knowledge and experience an invaluable resource. By setting a perfect example, he taught and educated me to become a scientist. Undoubtedly, he merits the designation "Doktorvater".

I am also very grateful to all my committee members, Tony Ciavarelli, Rudy Darken, John Hiles, Tom Lucas, and Kevin Squire for their help, support, and all the fruitful discussions. Specifically, I want to mention Tony Ciavarelli whose excitement about my research was always very reassuring, and John Hiles for sharing his very broad knowledge with me, and always pointing out connections of my work to closely and not so closely related fields of science.

Moreover, I am indebted to Nita Miller and Larry Shattuck for the constructive discussions and their support pertaining my eye-tracking experiment, and also for understanding that I considered their eye-tracker to be mine, Tim Chung for the excellent discussions and his valuable feedback throughout the last year, as well as Paul Evangelista for the great collaboration during the experiment phase of this work, and the extensive exchange of ideas thereafter.

I would also like to thank Danny McCue and Mike Guerrero for their help with programming the applications used for this work, Mike Dunhour for providing the artwork, and Steven Cyncewicz who edited the dissertation.

Very special thanks go to my sister Katja for her tremendous efforts and help during the final editing process of the dissertation.

Above all, I owe my deepest gratitude to my wife, Regina, for the huge sacrifice my work and my merely physical presence for almost four years involved, and for her infinite

support and encouragement. Lastly, I thank my sons, Tristan and Anselm, for their efforts of keeping me from working, and thus providing some of the most joyful moments.

Finally, I acknowledge that this research was partially funded by the U.S. Army TRADOC Analysis Center Monterey.

I. INTRODUCTION

A. THESIS STATEMENT

This research constructs a model of visual search for human targets in a military simulation which models human visual perception including eye movement behavior and attention deployment. The model input will be a computer-generated image representing the visualization of a situation that occurs during the run of a military simulation. The model will be based on the evaluation of low-level visual features, task-dependent cues derived from ground truth simulation data, and cognitive factors. The model output will be a map predicting the likelihood of eye fixations at each location of the map. This predictor map can subsequently be used to determine eye fixations.

B. PROBLEM STATEMENT

During the last twenty years, after the breakdown of the communist block the type of warfighting and the type of military conflicts have significantly changed. In the past, two opposing blocks that consisted of several countries would have fought against each other using weapons that employed high technology and posed high destruction capabilities. In these days the major weapon systems used were tanks, radar systems, airplanes, and missiles. These systems were in the focus of analysis to support decision making. The new type or style of warfare still relies on modern technology, but it became apparent in the current conflicts in Afghanistan and Iraq that technological superiority is not sufficient. In these wars the individual soldiers with their training, capabilities and performance, but also with their physical and cognitive limitations, play an increasing role. This means that the analytical tools used to inform decision-makers need to put more emphasis on the representation of individual soldiers. Foremost, human behavior representation has to be improved in order to gain adequate output from military simulations.

The modeling of target acquisition and detection has always been a major concern for military simulations. In the past, the capabilities of systems had been the focus of attention, now the capabilities and the performance of humans need attention. Military simulations are usually designed to serve a particular purpose. Based on this purpose the simulations intend to provide insights into different levels of abstraction or aggregation and different degrees of fidelity. Only a subset of these models, usually the ones which are less abstract and have a higher level of fidelity, consider individual soldiers at all. Some of these are JSAF, JCATS, CASTFOREM, OOS and COMBAT XXI. The stated simulation models have in common that they use the ACQUIRE algorithm for calculating the visual detection probabilities (`http://www.msrr.army.mil`). However, it has been shown that the ACQUIRE algorithm does not sufficiently reflect the performance of human observers (C. Darken & Jones, 2007). Additionally, in the experiments of C. Darken and Jones humans 'detected' false positives at a rate of around 10% in a visual search task. That means in more than 10% of the cases human observers thought they detected the target in a particular location where it actually was not. A similar finding has been reported by Henderson, Weeks, Jr., and Hollingworth (1999). In the study they performed, humans had to find pre-specified objects in line drawings of natural scenes, and the false positive rate was about 11%. This indicates that a good model of visual target detection cannot have detection probabilities as the only outcome, but has to produce false positive detections as well. ACQUIRE lacks this capability. One could argue that knowing the likelihood of false positives would allow for the extension of existing models by just generating the right rate of false positives with the usual mechanism of drawing random-numbers. However, this would still leave the question of where to place the false positives.

In addition to that, current target detection mechanisms in urban combat use the so-called 'windshield wiper' approach to determine which part of the scene the target detection mechanism is applied to (Harrington, 2009). This 'windshield wiper' approach assumes that the field of regard of a ground soldier is split into several fields of view. The field of regard is the sector around the soldier it is interested in and the field of view is a part of

this area, sized such that it can be considered to be visually processed as a whole. The field of views are adjacent and non-overlapping (Harrington, 2009). The target detection mechanism is applied to one field of view. It determines whether a target can be found in that area, and how long it takes to find the target or how long it takes to terminate the search within this field of view if no target is present. Then the next field of view becomes the active field of view and the search mechanism is applied there. This is done for one field of view after the next. When one end of the field of regard is reached the process starts from the other end again, hence the name 'windshield wiper'. This way of determining the locations to which the target detection mechanism is applied to is far from actual human behavior. This could be considerably improved by employing a model of likely fixation locations, which will be one of the contributions of this work.

Essentially, this research is concerned with the human eye movement behavior during target detection. This behavior is influenced by a variety of factors including human perceptual capabilities and external factors affecting the covert and overt deployment of visual attention. Human visual perception is mainly characterized by the receptive qualities of the retina. The fovea, which is the center of the retina, provides high visual acuity. This acuity decreases with higher eccentricity from the center. The field of high visual acuity is considered to subtend about 2° of visual angle whereas the complete visual field covers about 180°. The high acuity of the center is necessary for reliable object recognition. Outside of this high acuity area the received picture is very blurry. In order for humans to perceive the whole world around them with high acuity they have to perform eye movements. These movements allow humans to serially fixate objects in the visual field one after the other. The information derived from these serial fixations is then integrated into a coherent picture by the human brain. Additionally, visual attention is necessary to encode visual information into objects. The deployment of visual attention and eye movements are tightly coupled (Hoffman & Subramaniam, 1995) and are controlled by mechanisms which rely on perceived visual information, on task demands, and also on cognitive factors in order to determine the series of eye fixations and loci of attention allocation.

This research effort will generate a sophisticated model of eye movement behavior which will be capable of predicting likely fixation locations being examined by humans looking for a human enemy target in ground combat. The model needs to employ visual and cognitive factors as well as task demands that influence eye movements and the deployment of visual attention. These fixation locations can subsequently be used to apply appropriate target detection mechanisms at the fixation locations, and they are also the only locations at which false alarms are allowed to occur.

C. APPROACH

In order to form a basis for a target detection mechanism which is better aligned with actual human behavior and also incorporates false target generation, an eye movement model is to be created. Since eye movements are controlled reflexively based on visual scene features as well as through volitional components based on the actual task demands, both bottom-up as well as top-down influences need to be included into the desired eye movement model.

Unfortunately, it is unclear how top-down and bottom-up features are combined to yield the subsequent fixation location. Neurological research indicates that the two mechanisms are located in different areas of the brain. The frontal eye field is involved in the processing of the top-down mechanism. The bottom-up mechanism on the other hand is located in the superior colliculus. Finally, the information of the two mechanisms converges on the intermediate layer of the superior colliculus (R. M. Klein, 2004), which in turn drives the oculomotor system, which is the system that is responsible for the control of eye movements (Kandel, Schwartz, & Jessel, 2000). This shows that there are actually two mechanisms in the brain, which feed into one instance of eye movement control. However, it remains unclear how the control of the oculomotor system uses the incoming information. Estimating how each of these mechanisms contributes to the actual determination of the fixation site is the subject of this research.

In order to create a model of eye movements which includes bottom-up information as well as top-down information, it is first of all important to know how bottom-up and top-down information influence visual attention allocation. The results of previous research addressing this question are controversial (Bacon & Egeth, 1994; Theeuwes, 2004). The main concern of these studies was to determine whether bottom-up influences override task demands or vice versa. The results of Bacon and Egeth (1994) indicate that task demands override bottom-up influences, whereas Theeuwes (2004) finds the opposite. In addition to that, the experiments addressed attention allocation in general and not specifically eye movements. More recent studies also illuminated how the variation of target and distractor salience affect the influence of top-down and bottom-up information on search performance and eye movements (Born & Kerzel, 2008), but studies of this kind have been very rare so far. According to Born and Kerzel there has only been one other study examining how variations in target and distractor salience influence eye movements in a search task. However, both studies as well as the previously mentioned studies (Bacon & Egeth, 1994; Theeuwes, 2004) examined targets that were abstract shapes only and not real world targets. Therefore, it is questionable whether the results extrapolate to the search for a concrete real world target like a human enemy soldier.

In addition to that, in the experiments of the aforementioned studies, task-related information that has a meaning for the search task is not taken into account. Apparently, this is not possible, since the stimuli simply do not contain any such information. No location on the uniformly colored background contains any information content which would make an observer expect the target with higher or lower likelihood. The only task influence that can be incorporated is through the target, and top-down processing is solely engaged through pre-specifying the target features.

Eye-tracking data captured in previous experiments in which human participants searched for human target figures in realistic scenes (Wainwright, 2008) indicated on a qualitative basis that a substantial amount of eye fixations is directed towards locations where enemy ground soldiers could hide or blend in well with the environment. This indi-

cates that information which is meaningful for the search task is extracted from the scenes, cognitively processed, and used to inform the search in order to improve search performance. This means that an eye movement model that is used to improve target detection mechanisms needs to represent this type of information. Unfortunately, not a lot of information about that type of actual task influence, its implications for search performance, and how it affects eye movements can be found in the literature. The most informative research of this kind was looking at searches for objects, either abstract objects or real world objects, which were repeatedly presented at the same location of a particular scene. It was shown, that this repeated presentation of scene arrangements eventually improved search performance (Brockmole, Castelhano, & Henderson, 2006; Brockmole & Henderson, 2006). This means, that participants learned the co-occurrence of objects with scene locations and could therefore reduce search time. However, this does not answer the question whether meaningful scene content, which does not have to be explicitly learned over the course of an experiment, can be extracted from scenes and used to guide the eyes in search for a target.

A search experiment that includes eye movement recording of human participants is conducted in order to address a series of questions. First of all, possible interactions of top-down and bottom-up information are examined without assuming that one or the other would take precedence on attentional capture and drawing of the eyes. Also, the influence of a semantically relevant scene location is assessed. Lastly, the effects of variations of target properties, distractor properties, and properties of the relevant scene location on eye movements and search performance deserve scrutiny. The properties to be varied are eccentricity and salience.

In this experiment, the top-down information is represented by the target, which is a ground soldier in camouflage uniform, and by a semantically relevant scene location. This semantically relevant scene location is a doorway in the background wall, in front of which the target is presented. Based on observations of earlier experiments, which qualitatively showed that likely hiding locations receive a substantial amount of fixations, it is expected

that the doorway will be perceived as a hiding location and will be fixated by participants in the search for a human target. Examining how this semantically relevant scene location affects the eye movements will provide insights as to how much such a location contributes to the human eye fixation allocation mechanism.

The bottom-up influences are tested using a visually salient distractor object. This is an unfolded newspaper seemingly attached to the background wall. The main idea behind this design is to create stimuli that, although being rather simple, represent a real world scene and not an abstract search array. This real world scene would fit the specified task and would therefore provide actual task-dependent influences.

The knowledge gained in this experiment furthers the general understanding of eye movement in a search task, in particular in the search for a human target. This is necessary to gain information needed to inform the modeling of eye movements for target search.

As will be seen from the results of the experiment presented in the next chapter, the semantically relevant scene location actually draws the eyes in the search for the target. This information will be used for the created eye movement model.

The model will take a computer graphics-generated visual scene as input, and it consists of two major parts: a bottom-up part and a top-down part. The bottom-up part is a re-implementation of the salience-based visual attention model first described by Koch and Ullman (1985) and realized in an implementation of L. Itti et al. (1998). The term salience in this context means that some visual features stand out from the surrounding background and are thus salient. A salience map is an abstraction of an image. The values in the salience map determine how prominent a location in the image is, that is, how much it stands out from its surround. The top-down part on the other hand focuses on extracting semantically relevant information from a simulation environment, which yields a relevance map. This relevance map is a unique and novel development of this research.

The relevance map is derived from ground truth simulation data in several steps. First of all, a waypoint mesh is constructed which densely covers the simulation environment with waypoints. At each location in the simulation environment, which is reachable

by a human, a waypoint is placed by an automated process. For the desired viewpoint of an observer in this simulation environment three-dimensional scenes are rendered several times with one target figure being placed at one of the waypoints. The number of scenes rendered is equal to the number of waypoints visible from the viewpoint. Each time the scene is rendered, visibility information is computed for the target. This visibility information is stored in a three-dimensional data structure, the so called pixelbank at the respective (x,y,z) coordinate of that pixel on a per pixel basis. The visibility information computed includes contrast of the target with the background, the number of visible target pixels, and the fraction of visible target pixels over the total number of target pixels.

From the pixelbank, two two-dimensional top-down maps are computed. In one map, likely hiding locations are highlighted. These hiding locations are scene locations at which targets can take cover. This is indicated by the fraction of visible target pixels. Where this fraction is small, but not zero, targets are partially occluded and can therefore take cover behind objects, corners, doorways, or window frames. The second map highlights locations at which targets blend in well with their background. This is done using the contrast information. When targets have a low contrast they blend in well with the environment and are hard to detect. If the contrast is zero, targets are effectively indistinguishable from the background. These locations are of less interest as are the locations with very high contrasts. These two maps are additively fused into the final relevance map.

The relevance map, the re-implemented salience map, the original salience map, and a combined relevance/salience map are all compared to eye fixations collected from participants searching realistic scenes for enemy ground soldiers. This experiment uses realistic stimuli which are designed to represent scenes a ground soldier could encounter in urban combat or while conducting a foot patrol. A stimuli example can be seen in Figure 1. The targets in the scenes can be well-hidden as well as fully exposed and easy to spot, and the targets can assume one of four different postures.

The fixations collected from the eye-tracking experiment are fused into a fixation map, which is a binary map of the same width and height as the stimuli. Every pixel which

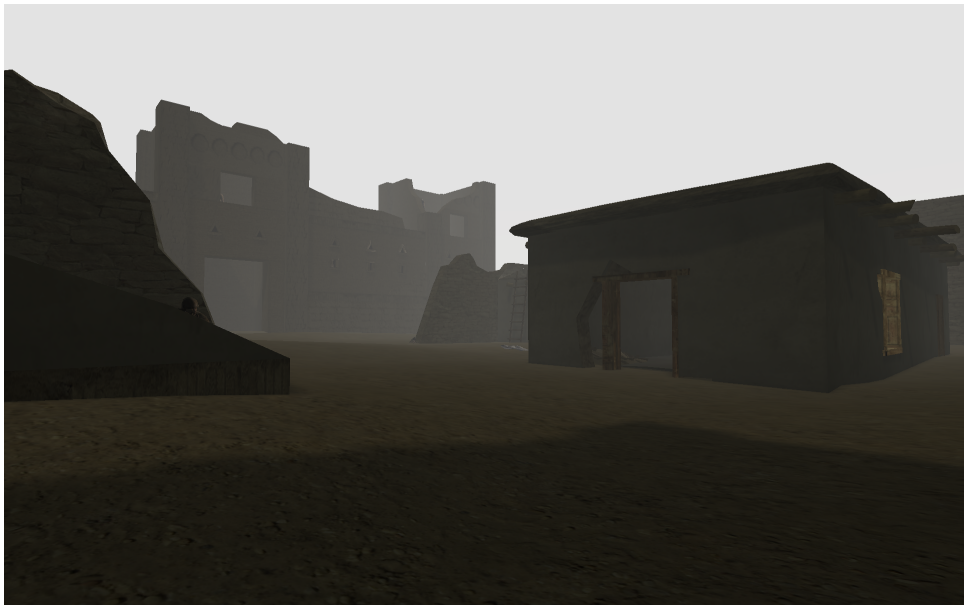

Figure 1: A stimulus example containing four targets. In the foreground on the left, there is a target kneeling behind the little wall. In the background, one target is standing in the window, another target is standing behind the wall on the right tower. In the center of the scene, there is a target on the ground behind the wall left of the ladder.

received a fixation is marked with a one, and all pixels that did not receive a single fixation are marked with a zero. These fixation maps are used to assess how well the salience maps, the relevance map, and the combined map predict fixations. This is done by treating the maps as predictors or classifiers. The ones and zeros in the fixation map represent positive and negative instances of fixations and the predictor maps are evaluated using the area under the receiver operating characteristic curve of each of the predictor maps. The area under the curve (AUC) is known to be equivalent to a Wilcoxon rank sum test, and it therefore tells how likely it is that positive and negative instances are ranked correctly (Hanley & McNeil, 1982).

D. CONTRIBUTIONS

The contributions of this work are of two types. Firstly, the visual search experiment with real world targets and a realistic task provides new insights into the interaction of

visual salience and task-dependent information in general, and especially in the area of target detection. Secondly, and possibly even more importantly, a novel technique for representing top-down target search information is described and shown to significantly improve on previous work in predicting eye movements.

The first important finding of experiment 1 is that neither top-down nor bottom-up information overrides the other. Although the task demands have a stronger influence on the eye movements, a visually salient distractor can always draw the eyes of an observer. This is a very important finding that differs from earlier findings which attributed overriding capabilities with respect to attentional capture to either top-down or bottom-up factors (Bacon & Egeth, 1994; Theeuwes, 2004; Einhäuser, Rutishauser, & Koch, 2008).

In fact, the strength of attentional capture strongly depends on the target and distractor properties. Most interestingly, the attentional capture potential of the distractor does not increase with increasing salience. It only increases up to a point and then the distracting potential declines. This is a rather surprising result since it is generally assumed that higher salience leads to higher attentional capture. One of the most prominent computational models of visual attention heavily relies on this assumption (L. Itti et al., 1998). The findings here show that this assumption is misguided.

The finding that search performance and eye movement behavior are influenced by target and distractor properties significantly extends the earlier observations of this effect in the absence of a concrete task (Born & Kerzel, 2008). The research presented here shows that the same effect is present in case of specific task instructions as well. In addition to that, the effect was observed for a variety of eye movement metrics whereas the earlier research could show this effect for one metric only.

The most important finding of the experiment is that humans use semantically relevant scene locations to improve their search performance if they cannot spot the target easily. These meaningful scene locations also guide eye movements. The particular type of meaningful scene locations studied in this work is hiding locations. Hiding locations

are cognitively processed as such and are subsequently used to inform the search task, presumably with the goal of speeding up the search process.

Another contribution of the experiment is its unique design, especially with respect to the employed stimuli. Roughly resembling stimuli following the visual search paradigm, they still resemble real scenes (for example, photographic images). They are therefore assuming an intermediate position between abstract search arrays and realistic scenes. The design allows experimental control of targets, distractors and hiding locations, which for realistic scenes is virtually impossible. Once again, with the hiding location the stimuli contain a semantically relevant scene location. Abstract search arrays do not allow the inclusion of elements with meaning for the search task.

The most important contribution of this work is a new kind of eye fixation predictor. The idea of this predictor, termed the relevance map, is similar to the idea of using a salience map, which predicts eye fixations based on visually salient scene elements. However, the novelty behind the relevance map is that it represents scene locations that are semantically relevant for the search task. The creation of this map is inspired by the results of the search experiment, which showed that meaningful locations are used to improve the search process and guide eye movements. Therefore, these meaningful locations need to be considered by a model of eye movements.

Extracting meaningful information for a search task from the ground truth data of a simulation is another novelty provided by this research. This information is used to construct the relevance map. There is no other research capturing scene information in this way, and there is no other research capturing scene information of this kind. With this novel approach high level task-dependent information is accessed with higher accuracy and better quality than could be done with the more traditional computer vision approaches. Although the computer vision approach follows the biological facts of human information retrieval more closely, it poses a lot of unsolved challenges, especially relevant to the generation of higher level knowledge. The approach used in this work on the other hand pulls ground

truth information from the simulation such that using the cumbersome step of resembling the human sensing mechanism becomes superfluous.

Comparison of the relevance map with eye fixations show that the relevance map predicts eye fixations very well. Even better prediction performance is achieved by combining the relevance map with a salience map. This combined map takes task information as well as visual information into account. These maps provide a foundation for future eye movement and visual attention models that are based on visual scene information as well as on meaningful scene content with task relevance.

II. BACKGROUND AND RELATED WORK

In order to model human behavior as described in the previous section, a variety of research areas need to be taken into account. These areas can be subdivided into two major components: human aspects and computational aspects. The human aspects include physiological, psychological, and psychophysical elements such as visual attention, eye movements, sensation, and perception as well as neurological signal processing in the human brain. The computational side consists of existing computational representations of theories in the aforementioned areas and also mathematical and information technology tools necessary to build computational models. Throughout this chapter, no attempt is made to assign the described topics into one specific class since each topic usually covers more than one area.

First of all, the theory of visual attention has to be considered. Like attention in general, visual attention poses limits on the processing of signals and information coming into the human brain. These limits seem to be a crucial part of scene perception mechanisms which need to be understood. It will be shown that attention and eye movements are intimately related and thus eye movement research will also be looked at. Of special interest is the work which sheds light onto the relationship of visual attention and eye movements. Since this work has a focus on target detection, it is important to draw information from previous research in classical visual search and object search in naturalistic scenes, which is closely related to eye movements in scene perception.

A. VISUAL ATTENTION AND EYE MOVEMENTS

The human sensory system receives a vast amount of information. The data rate is estimated to be in the range of 10^7 to 10^8 bits per second at the optical nerve alone. This amount of data needs to be filtered such that the brain is able to cognitively process the information. The human nervous system employs an attention mechanism which effec-

tively filters the incoming information. Thus, information is processed serially on a subset of the available data instead of parallel due to computational limitations (L. Itti & Koch, 2001a). This requires the use of an intelligent mechanism to determine which information is attended. Although this statement holds for attention in general, visual attention is considered first. Visual attention is usually described with the spotlight metaphor indicating that only some highlighted part of the visual field is active for higher-level processing. In order to get a complete picture of the surrounding world, the spotlight will illuminate different parts of the field of view in a sequential manner. It has been believed that the spotlight has to cover a larger area in order to pay attention to multiple objects in different locations. However, recent research indicates that this metaphor does not hold true in every detail. The area of attention does not necessarily have to be continuous, which means there can be more than one spotlight (Cavanagh, 2004). No matter what the attention area actually looks like, it has to be redirected from time to time in order to cover all essential elements of the current field of view. This control of visual orienting is performed by two mechanisms: exogenous control and endogenous control of attention. In addition to these two different types of control, the deployment of visual attention comes in two flavors: overt shifts in attention and covert shifts in attention. The following two subsections will take a closer look at these four mechanisms.

1. Exogenous and Endogenous Control of Attention

Exogenous control is influenced by scene features like colors, patterns, or textures and is reflexive in nature whereas endogenous control is voluntarily and influenced by task demands, interest, or other cognitive factors (L. Itti & Koch, 2001a; Frintrop, 2006). L. Itti and Koch (2001a), as well as (Frintrop, 2006), call the feature-based mechanism bottom-up control and the voluntary mechanism top-down control. R. M. Klein (2004) uses the terms exogenous or reflexive control instead of bottom-up and endogenous or voluntary control for the top-down mechanism.

As already mentioned, the exogenous control of attention relies on features within the visual field of the observer. The set of these features is called sensory stimulation

pattern by R. Klein, Kingstone, and Pontefract (1992) and lies outside of the observer, thus the name exogenous. These features, however, do not draw attention just for being there but because they stand out from the background (Desimone & Duncan, 1995). It has been shown in search tasks that an object that is different in one feature dimension (e.g., color) from the other objects present will "pop-out" and automatically draw attention (L. Itti & Koch, 2001a), thus the name reflexive control. Several feature dimensions that are able to elicit a pop-out have been identified; these include color, orientation, spatial frequency, brightness, and orientation of movement (Treisman & Gelade, 1980). Often, this standing out of objects from the background is called saliency (Koch & Ullman, 1985; Itti, 2003) and several computational models for visual attention have been built around this idea of saliency (refer to Section II.C on page 33 for more details). It is assumed that saliency is computed pre-attentively across the entire visual field (L. Itti & Koch, 2001a; Koch & Ullman, 1985). Recently, this idea received support from neurological experiments. The results from the experiments indicate that a saliency map for basic features is stored in area V1 of the primary visual cortex (Zhaoping & Dayan, 2006).

As most people know from their own experience, humans do not deploy their attention based solely on the salience of items in their environment. When humans need to perform a certain task, like figuring out what time it is, they direct their gaze right away to the spot on the wall where they know the clock is located. Presumably, this direction of attention is facilitated through pre-attentively available information, but it is not purely reflexive, but also strongly volitional. The influence of volitional control on eye movements was observed in one of the earliest eye movement experiments. Yarbus (1967) found out that humans who looked at an image exhibited completely different gaze patterns when asked to answer different questions regarding the scene on the image. Additionally, the results of Yarbus show that there are high similarities both within subjects and across subjects when free-viewing the same picture. The within-subject similarities are the result of several exposures to the same image with an interval of 1 to 2 days in between exposures. However, the within-subject similarities are higher than the across-subject similarities. Al-

together this indicates that attention is directed to different parts of the image, depending on the task an observer has to perform.

Noton and Stark conclude from their experiments that eye fixations during object recognition are purely volitional and that there is no influence of low-level features. They argue that influences of low-level features would elicit the same eye movements for every observer (Noton & Stark, 1971). At first glance, this seems to be a contradiction to Yarbus' findings, since his results indicate similar gaze patterns across subjects for the same task. Now if the same task elicits the same eye movements for different observers, then this would mean that not having the same eye-patterns implies that the task was not the same and thus the argumentation of Noton and Stark would be invalid. This, however, is a fallacy. Although they show high similarities, the gaze patterns observed by Yarbus are actually different for different observers. Yarbus assumes that this is due to the different ways in which people think. On the other hand, for reflexive reactions in general one would expect to see only slight differences across subjects, if at all. Thus, one would expect to see almost identical scan patterns across subjects if they were generated by a reflexive mechanism. Since this is not the case, the argumentation of Noton and Stark appears to be valid. Unfortunately, the data of Noton and Stark as well as the data of Yarbus are evaluated on a qualitative basis only. It would be interesting to use quantitative measures for further evaluation. However, the qualitative analyses provide sufficient information to conclude that a task-dependent or top-down mechanism influences eye movements.

Similarly, R. Klein et al. (1992) point out that eye movements are controlled strategically to locations of the scene that contain relevant information. However, they do not provide any idea as how the brain determines which regions contain the important information. This usually is the concern of psychological research interested in scene perception, target search and object recognition (Rayner & Pollatsek, 1992; Horowitz & Wolfe, 1998; Oliva, Torralba, Castelhano, & Henderson, 2003). This research tends to use the term 'top-down', whereas in physiological and neurological research the term 'endogenous' is prevalent, (e.g. R. Klein et al., 1992). This latter research usually uses cueing by visual

or auditory stimuli to elicit endogenously controlled attention shifts and measures reaction times and accuracies of task performance. With this approach it has been shown that endogenous direction of attention is rather slow compared to exogenous control which works considerably faster (R. M. Klein, 2004). It has also been shown that endogenous control can be affected by memory load. These and other findings led to the conclusion that endogenous and exogenous control of attention is performed through isolable subsystems (R. M. Klein, 2004).

Clearly, these two systems do not operate separately from another but in parallel. However, still one of the two or a combination of the two control mechanisms will determine where the attention will be directed to. Thus, there must be some interactions going on between the two mechanisms. The essential question is: how are competing shifts of attention, produced by the two different subsystems mediated? No general answer has been found yet. However, R. M. Klein (2004) as well as Theeuwes, Kramer, Hahn, and Irwin (1998) observed that exogenous orienting overrules endogenous orienting when two stimuli triggering the respective subsystems occur simultaneously. Theeuwes et al. also report that this overruling does not take part when the endogenous cue is presented well before (200 ms) the onset of the exogenous stimuli. In addition to that, there is a dispute whether task demands can override stimulus-driven attentional capture. Employing the visual search paradigm, Bacon and Egeth (1994) find that task demands can override stimulus-driven attentional capture. Using a very similar experimental design, Theeuwes (2004) shows that attentional capture can only be overridden if the distractor saliency is not high enough. When the distractor salience is raised, task demand cannot revert attentional capture by the salient distractor.

Very interestingly, the two different mechanisms for exogenous control and endogenous control respectively differ in the functions they are able to perform. In their feature integration theory, Treisman and Gelade (1980) postulate that attention is a necessary factor which enables the brain to integrate extracted features into a unitary object. However,

R. M. Klein (2004) reports that exogenously deployed attention is capable of performing feature integration but not endogenous orienting.

In his research, R. M. Klein also looks at differences in endogenous and exogenous control depending upon whether they are employed for overt or covert visual attention. This level of detail is not relevant for the present research, but the general distinction between overt and covert attention is important. In contrast, there are speculations, that objects might be formed before attention is directed to them. Subsequently, eye movements could possibly be guided through object saliency rather than through feature saliency (Einhäuser, Spain, & Perona, 2008).

2. Overt and Covert Attention

The main difference between overt and covert attention is the involvement of eye movements. Covert attention is the deployment of attention without eye movement, whereas overt attention is associated with eye fixation. The movement of the eye indicates a shift in overt attention to the new eye position, whereas the eyes keep fixating on the same location while covert attention shifts are performed. The overt deployment of attention serves the important necessity of directing the gaze to the most important location of the visual scene. This allows the fovea, the most sensitive part of the human retina due to the high density of photoreceptors, to be positioned on the important parts of the scene in order to gain information with the highest possible acuity.

The relationship between overt attention and eye movements has been shown by Hoffman and Subramaniam (1995). In a set of experiments they showed that subjects are not able to deploy attention to one point and move the eyes to another point at the same time. This demonstrates that eye movement is coupled with overt attention deployment. Additionally, their results indicate that eye movement is faster when the new fixation location is known in advance, suggesting that covert attention deployment to the new location might precede eye movement. Despite or maybe because of this relationship between eye movements and attention shifts, one needs to be careful not to treat them as equivalent. The location of attention does not equal the location of fixation; it is the shift of attention which

is coupled with the shift of fixation. After the eye movement, it is only known that attention is deployed to this particular fixation location for some amount of time. Thus, at any moment in time, one cannot infer the location of attention from the current eye position.

On the other hand, according to the "active vision" perspective, covert attention allocation is just a supplement to the actual movements of the eyes. It is claimed that no covertly serial scanning takes place during fixations, and that overt scanning is the normal way of attentional shifts (Findlay & Gilchrist, 2001).

B. EYE MOVEMENTS AND SCENE PERCEPTION

According to Rayner and Pollatsek (1992), we are far from understanding exactly how humans extract information from a panoramic scene. Since then, a lot of light has been shed on the mechanisms of scene perception, but still many open questions remain. This research mainly focuses on using already available information, although some unanswered questions will need to be addressed, and some of the available answers need to be illuminated from a new angle. The scene perception research tries to understand the relationship between cognitive mechanisms of scene perception and the related eye movements with the final goal of developing conceptual and sometimes computational models as well. In contrast, the main interest of this research is not to develop a generally applicable model of visual attention for scene perception. Instead, the goal is to generate a computational model which requires additional and slightly different information, mostly because the desired model will cover a specific niche of scene understanding involving perception-limiting conditions and usually hidden targets. The questions that need to be answered in order to create the model are not directly addressed by other researchers. Nevertheless, a lot of information is already available and several interesting lines of research are related to this work or provide background information.

One reason for eye movements during scene perception is the decreasing acuity of human vision with higher eccentricity from fixation (Henderson, 2003; Rayner & Pollatsek, 1992; Rayner, 1998). The central part of the retina, which covers the fixation point and the

2°of visual angle of the field of view next to fixation, is called the fovea. The acuity is due to the very high density of photoreceptors. Rayner (1998) calls the part of the retina adjacent to the fovea parafovea and defines that its field of view extends 5° of visual angle on either side of fixation. Its acuity is less than the acuity of the fovea but higher than the one of the periphery, the remaining field of view. It is important to note that the retinal acuity does not degrade abruptly at the respective boundaries; it rather decreases continuously with higher eccentricity from the fixation. This is already true even for the fovea itself (Rayner & Pollatsek, 1992). However, the rough classification provided by this terminology eases the descriptions and explanations of eye movement experiments in scene perception tasks, and this work adopts this nomenclature. Another reason for eye movements is the deployment of attention to different parts of the scene together with the eyes. As already described earlier (Section II.A.2 on page 18), visual attention is coupled with eye movements and is also necessary for object encoding. However, overall it is unclear what signals tell the eyes where to move (Rayner & Pollatsek, 1992).

The question at the heart of scene perception research is: What is the eye movement behavior of humans observing a scene and what does this behavior imply for the underlying control mechanisms in conjunction with the cognitive processing of the scene?

Answering this question first of all requires understanding the meaning of "scene perception" in this context. Rayner and Pollatsek (1992) define scene perception as the identification of a scene setting and the most important objects it contains. This definition subsumes object detection in a given scene since this target object has to be considered one of the "important objects" if not the most important object.

As already stated, the acuity of the human visual field degrades rather rapidly from the center to the periphery. Introspection, however, reveals that humans are able to acquire a great deal of information from outside of fixation. In the following subsection the differences between information derived centrally and peripherally are established.

1. Foveal Versus Extra-foveal Processing

In order to understand the importance of foveal and extra-foveal information capture and processing in scene viewing two different experimental paradigms have been developed. In the first paradigm, the so-called moving window paradigm, initially developed for reading experiments, a moving window is created around fixation. This requires a high acuity eye-tracker with high temporal resolution. Based on eye tracking data, a mask is superimposed on the viewed scene, leaving a window only around the fixation point. The window size can be adjusted according to the needs of a specific experiment. In the second paradigm observers view a scene for a limited time while eye movements are recorded. After participants have observed the image they have to answer questions about the objects in the scene. After comparing the results of the answers with the eye fixations, it turned out that subjects had no information about objects further than 2.6°of visual angle from any fixation point. This result raises further questions since it is not clear whether the effects are due to acuity, attention, or maybe both. All objects which had not been fixated but still encoded must have been processed by extra-foveal vision. Additionally, from the experiments of Hoffman and Subramaniam (1995) it is known that the deployment of covert attention precedes eye movements and covert attention is related to extra-foveal vision per se. Altogether, this raises the question of what kind of information can be extracted from extra-foveal vision which first guides covert attention and subsequently elicits eye movements.

Henderson, McClure, Pierce, and Schrock (1997) examined the importance of foveal and extra-foveal vision for object identification by using an artificial scotoma paradigm. In these experiments either the fixation location or a location with some distance off of fixation were masked either with a gray patch or a placeholder representing the scotoma. As a control condition, one-third of the trials was ran without any masking. Although the identification performance was worse in the scotoma conditions, it was still way above chance, thus indicating that humans can extract a considerable amount of information extra-foveally. The better performance of the unmasked condition, however, shows the importance of the

high acuity information gained by the fovea. Interestingly, in the center-scotoma condition participants tended to foveate regions in between objects, which is in contrast to results of other experiments (Henderson et al., 1999). This suggests that humans prefer gaining information with the highest possible accuracy. Another explanation would be that the fovea is directed to blank space in order to facilitate covert attention direction to the object that is to be encoded. This idea would be in accordance with the crowding theory of Cavanagh (2004). According to this theory, attention can be deployed to larger areas, but from a set of nearby objects in the extra-foveal visual field only partial information can be derived. Another interesting finding of Henderson et al. (1997) can be derived from the unusual eye movement behavior in the scotoma condition. The observed difference from the baseline shows that the humans were able to adapt their behavior to compensate for the induced handicap. Furthermore, looking at the performance results, it becomes obvious that the compensation was successful and that subjects could derive significant object information from the parafovea. This has already been reported by Rayner and Pollatsek (1992), but they also assume that information can similarly be extracted even from the periphery.

Another fact derived from the artificial scotoma experiment of Henderson et al. (1997) points out the importance of the information gained from the periphery. When the artificial scotoma was placed with an offset to the fixation location, i.e., extra-foveal vision has been limited, the processing time of fixated objects increased. Apparently, object encoding usually already takes place when overt attention is directed to the object before its foveation (Henderson et al., 1997). Saida and Ikeda (1979) derived similar results from a moving window paradigm. They varied the size of a rectangular window and observed how this affected picture recognition. With a window size of about $3.3° \times 3.3°$of visual angle, performance in picture recognition decreased dramatically. It improved with larger window size and finally became asymptotic (Rayner & Pollatsek, 1992). Similarly, Henderson et al. (1997) report that the visual system can gain object information from up to $10°$ from fixation, but only if the peripheral field of view is otherwise empty, which matches with the crowding theory of Cavanagh (2004).

2. General Eye Movement Patterns

This section looks at measures of eye movement behavior and general observations regarding eye movement patterns which have been found in previous research. The patterns of eye movements give some indication about the range of possible eye movements, which needs to be taken into account when generating an eye movement model. More importantly, these patterns allow inferences of the processing that is going on when saccades are programmed as well as to what information is extracted from scenes and the subsequent processing of this information.

Several different measures have been used in the past, but the ones which recur most often are saccade length and fixation time (e.g. Henderson et al., 1999; Rayner & Pollatsek, 1992; Mannan, Ruddock, & Wooding, 1997; Torralba, Oliva, Castelhano, & Henderson, 2006). A very extensive set of measures is used by Henderson et al. (1999), who attempted to understand how scene content and semantics influences eye movement behavior. They define three categories of measures: measures of extra-foveal semantic analysis which include probability of immediate target fixation, number of fixations to target and amplitude of initial saccade to target; measures of fixation density including proportion fixated and numbers of entries into target region; and measures of processing time: first pass gaze duration, first pass gaze fixation count, average first pass fixation duration, second pass gaze duration, total fixation count, and average fixation duration. However, the following description focuses on saccade lengths and fixation times. Rayner and Pollatsek (1992) report saccades lengths of 2° of visual angle, 4° of visual angle and 3° of visual angle and fixation times of 225 ms, 330 ms and 275 ms being typical for reading, scene viewing and visual search, respectively. They also note that there is a relatively high intra-subject variability of saccade lengths ranging from 2°of visual angle up to 6 °of visual angle. In addition to that, Henderson et al. (1999) observed differences for saccade length depending on the task and the saccade targets (objects). The objects could be consistent or inconsistent with the scene context. An object that is consistent with the scene context would be fit into a particular scene, for example a microscope in a laboratory. An inconsistent object

would discord with the scene context such as a microscope in a bar. The task of the participants was either memorizing the scene content or searching for an object. The average saccade lengths in the memorizing task was 2.86° of visual angle for inconsistent objects versus 3.21° of visual angle for the consistent objects whereas the average saccade length in the object search task was 3.49° of visual angle for inconsistent objects versus 3.86° of visual angle for consistent objects. The saccade lengths were only reported for saccades going to pre-specified objects and no data was reported for all other saccades, so no information is available about the range of saccade lengths in this experiment. This means no comparison is possible to the intra-subject variability provided by Rayner and Pollatsek (1992). In opposition to these findings some researchers have observed long saccades early in scene viewing and assumed these were exploratory saccades (Rayner & Pollatsek, 1992). Similarly, Henderson et al. (1999) review experiment results in which observers showed saccade lengths at around 7° of visual angle. Henderson et al. attribute these differences to object density in the scene. They assume that sparsely populated scenes elicit larger saccades than usually observed. Altogether these findings provide converging evidence for an upper bound for average saccade lengths in scene viewing. This upper bound is usually around three to four degrees of visual angle in scene viewing, but can be higher in sparsely populated scenes.

Another important aspect in eye movement patterns is the similarity of fixation locations on the same or possibly on different scenes across observers, and across repeated viewings of one scene by the same subject. To calculate these similarities Mannan et al. (1997) suggest the following index of similarity:

$$I_S = 100\left[1 - \frac{D}{D_r}\right] \tag{1}$$

where

$$D^2 = \frac{n_1 \sum_{j=1}^{n_2} d_{2j}^2 + n_2 \sum_{i=1}^{n_1} d_{1i}^2}{2n_1 n_2 (a^2 + b^2)} \tag{2}$$

and $n_1$and n_2 are the number of fixations in the fixation patterns, d_{1i} and d_{2j} are the distances for fixation i or j in set 1 or 2 respectively to its nearest neighbor in the other set and a and b are the width and the height of the image. D_r is calculated like D but the fixation patterns are two different random patterns with the same number of "fixations" n_1 and n_2 as used in the calculation of D.

However, as noted by Henderson, Brockmole, Castelhano, and Mack (2007) the distance metrics used for the similarity index could possibly compare all fixations of one pattern with a single fixation of another, which is very close, whereas all other fixations are quite far away. Thus they suggest that a one-to-one assignment of fixation locations of one pattern to fixation locations of the other pattern is performed first, then D is calculated as follows:

$$D = \frac{1}{n} \sum_{j=1}^{n} p_j^2 \tag{3}$$

where n is the number of fixations and the values of p_j are the distances between the unique pairs of fixation-locations (one location of each fixation pattern assigned to each other).

This calculation requires that the patterns to be compared have the same number of fixations, which is not generally true. Unfortunately, Henderson et al. (2007) do not provide a solution for that. One could easily assign a unique location to every location of the pattern with less cardinality and assign the remaining locations of the larger pattern to its nearest neighbor.

In their experiment Mannan et al. (1997) found out that the similarity of eye movement patterns on the same scene for different observers decreases over time. Also, the similarity of eye movement for one image is about the same across subjects and within subject. The within-subject measure was taken by presenting the same scene twice with 24 hours in between viewings (Mannan et al., 1997). This means that one observer viewing the same image twice showed different eye movement patterns each time.

The similarity index as defined above does not take into account the order of fixation. In fact, this has also been examined by Mannan et al. (1997). Even for any two subsequent fixations in a fixation sequence, they could not observe any repetition of fix-

ation orders across subjects. After having looked at different eye movement patterns, the following subsection examines how these patterns come about.

3. Top-down Versus Bottom-up Influences

It is pretty much agreed upon that eye movements are mainly influenced by visual scene features (bottom-up) and task-dependent requirements (top-down) (Henderson, 2003; Henderson et al., 2007; Itti, 2003; L. Itti & Koch, 2001a; Oliva et al., 2003; Rayner & Pollatsek, 1992; Yarbus, 1967). What is not clear yet at all is how the bottom-up and top-down controls interact with each other.

Rayner and Pollatsek (1992) assume that the eyes are mostly driven by lower level information. This is to some extent supported by the research of Mannan et al. (1997) who compared fixation locations with the locations of low-level scene features such as edges, contrast as well as brightness maxima and brightness minima. Their results show that some areas of the scene get higher numbers of fixations, but only locations with higher edge density or higher contrast level attracted fixations. Although this indicates that visual features play a role it does not prove that these features are the determining factor for gaze control. Especially considering the level of similarity between locations containing prominent low level features and fixation sites, which has not been very high, higher level features cannot be ruled out as factors contributing to eye movement control. In fact, Rayner and Pollatsek (1992) do not exclude higher level information as contributors and also acknowledge them as being necessary and important. However, they speculate that it would make sense if a simple process for eye movement control was used since such a process would require fewer resources from a "central capacity". The central capacity could then be used exclusively for object identification, which is considered a resource-expensive activity (Rayner & Pollatsek, 1992). Still, this disregards the possibility of covertly deployed attention before eye fixation which still would use the central capacity but does not move the eyes yet because attention movement is faster than eye movement.

In contrast to Rayner and Pollatsek (1992), Henderson (2003) suggests that higher level and lower level features need to be adequately combined. At the same time, he puts

the greater emphasis on the higher level based control he calls knowledge driven. The visual feature based gaze control is called stimulus-based. Henderson subsumes two types of established models into this category. The first type is the so-called scene statistics approach which looks at a range of image properties at fixation locations, and the second type is the saliency-based approach which determines regions differing substantially from their background. According to Henderson both do not allow a causal link between image properties and fixation locations and thus fall short. Also, the correlation of fixations and salient locations decreases when meaningful scenes are viewed and knowledge-driven gaze allocation increases over the course of viewing, thus modulating or even replacing saliency.

The category of knowledge-driven gaze control comprises three types: episodic scene knowledge, scene schema knowledge, and task related knowledge. The task-related knowledge is usually called top-down control. It is based on a task-specific control strategy or control policy. For tasks like driving a car or even more flying an airplane, specific, well trained gaze patterns are employed which are necessary to fulfill the specific task. The scene schema knowledge stores information about specific scenes such as spatial layouts and typical objects associated with a scene schema. This knowledge allows an observer to draw inferences from the scene of likely locations of objects, which possibly influence eye movements, for example, in a search task. In other tasks like driving a car this knowledge could elicit eye movements to scene locations important for specific situations. Say a driver is approaching an intersection; he or she will scan for traffic lights and traffic signs which have specific locations. Knowing in which country the driver is operating the car will guide the eyes to different locations based on his or her experience as to where the essential objects are usually located. The episodic scene knowledge is further subdivided into short-term episodic scene knowledge and long-term episodic scene knowledge. The former refers to knowledge derived from a scene over the course of viewing. The long term knowledge is accumulated over time when viewing a particular scene over and over again (Henderson, 2003).

The fact that Henderson considers episodic scene knowledge is very interesting because usually memory is not subsumed under top-down influences. However, recent findings indicate the relevance of memory for scene viewing (e.g., Hollingworth & Henderson, 2002; Hollingworth, 2006). They found out that a lot of information about scene content is retained over the course of scene viewing as well as for a longer period after exposure to the scene. This memorized information might influence eye movement behavior.

4. Global Versus Local Information

In this subsection local and global information will be contrasted and subsequently related to the concept of scene gist. Gist means the general context or category of a scene. Psychophysical experiments have shown that scene gist can be extracted from an image within a very limited amount of time. In tachistoscopic experiments, Biedermann, Mezzanotte, and Rabinowitz (1982) showed a series of images to observers. Each of the images was presented for 150 ms and the next image was shown immediately afterwards. This timeframe is so short, it does not allow for a single saccade, but still humans are capable of noticing whether a scene of a certain context has been among the presented images. Rayner and Pollatsek (1992) conclude that some kind of global, extra-foveally perceived information over the complete field of view must be used to generate the scene gist perceptually. This global information is seen in contrast to local, foveally perceived scene content. It is called global because the information can be extracted from a mechanism that does not require scanning of a scene; possibly this mechanism is attention free.

Schyns and Oliva (1994) explored the influences of scene gist on scene perception by looking at coarse and fine visual information. According to them, the classic approaches to scene perception assume that local information such as edges, shading, or motion is extracted from scenes and gradually combined to form objects and finally recognize scenes. This approach, however, would require people to perform several eye movements in order to encode a number of objects from which the scene gist could be derived. However, as described earlier, humans are capable of extracting the gist of a scene in a single glimpse, which indicates that the aforementioned scene perception theory is at least incomplete.

Schyns and Oliva (1994) propose another idea of how humans derive the gist of a scene. They claim that scene categories have a particular, inherent spatial layout of essential elements, and thus exhibit a global organization which is supposedly sufficient for fast scene gist recognition. Similarly, Brockmole et al. (2006) claim that the specific settings of naturalistic scenes are characterized by spatial arrangements of scene-typical objects even if the objects can be moved.

Schyns and Oliva (1994) generated high frequency filtered and low frequency filtered versions of one image. The high frequency filtering preserves the fine details, edges and boundaries from the original, whereas the low-frequency filtering only preserves coarse blobs. Schyns and Oliva (1994) assume that the spatial arrangement of these coarse blobs represents the global information associated with the scene category. In their experiment they showed that the relationship between scene gist and global information is stronger than the relationship between scene gist and local information. However, humans apparently do not rely on global features exclusively, even for exposure times as brief as 45 ms. This is in accordance with the statement from Rayner and Pollatsek (1992) that global information and locally extracted details interact. Furthermore, they assume that global information is extracted continually and not only at the beginning of scene viewing. This is a very reasonable assumption considering that the human visual system usually does not encounter still images but that humans are situated in their environment. Unfortunately, there are two competing systems again. Like Schyns and Oliva (1994), Brockmole et al. (2006) believe that scene identity is established at a global level. Furthermore, they assume that the scene identity drives attention to task-relevant scene regions. They speculate that scene identity and the underlying global information might guide observers to target positions which have previously been encountered in similar scenes (Brockmole et al., 2006). This theory is supported by evidence established by Brockmole et al. (2006) as well as by Castelhano and Henderson (2007).

Brockmole et al. (2006) presented a series of images to observers containing a prespecified target. A specific scene out of a large set was presented repeatedly to trigger

a learning effect. However, this scene had been modified in two different ways. Either some local context or global context was preserved. Assume a target was located on a coffee table in a living room. Maintaining the global context preserved the general room setting but changed the coffee table whereas maintaining local context put the coffee table of the original scene in a different room. The results of Brockmole et al. revealed that the greatest learning effect occurred in the condition which maintained global and local context, that is for repeated presentation of the exact same stimulus. Additionally, very good learning effects could be observed in the condition which maintained global context. A learning effect could also be observed for the maintained local context, but in this case the benefit was significantly inferior to the other conditions. This indicates that the information retained in memory is most likely some relationship between the target location and global scene features. This is in accordance with Rayner and Pollatsek (1992), who state that scene context can modulate object identification.

Where Brockmole et al. (2006) showed the influence of global information on target search, Castelhano and Henderson (2007) examined how the presumably global information derived from an initial glimpse influences eye movements in object search. They asked subjects to find a specified target in naturalistic scenes using a moving window paradigm. This moving window was placed at the eye position of the observer. An eye-tracking device captured the eye position and the window immediately followed the eye to the new location such that the fixated site could be examined but the rest of the image was masked. In the study of Castelhano and Henderson the disc-shaped window had a diameter of 2° of visual angle. Right before exposure to the masked stimulus, participants were shown a brief preview for 250 ms. This preview was either the scene which was shown later, another unrelated scene, or a meaningless scene with some local visual features such as colors, edges or lightning. The study showed that the identical preview condition was superior to the two other conditions in four different measures (response time, latency until first fixation on target, number of fixations, and ratio of saccade path-length taken to shortest possible distance from initial fixation to target location). The remaining two conditions

did not show any statistical significant difference in any of the four measures. This shows that the initial preview facilitates target detection and influences eye movement patterns but it does not show that this is due to global features. However, the results of Brockmole et al. (2006) and Schyns and Oliva (1994) indicate that humans extract mostly global information when exposed to images very briefly. This suggests that the preview benefit observed by Castelhano and Henderson (2007) is the result of global information derived from the initial glimpse.

It is also important to notice that in the study of Castelhano and Henderson global and local information must have interacted. In their experiment no global information was available during scene viewing due to the mask around the window at the fixation site. However, in the identical preview condition observers could capture the global information from the preview, whereas in the meaningless preview condition no global information was available at any time; this resulted in decreased performance in the meaningless preview condition compared to the identical preview condition. This means that the local information alone was inferior to the local information combined with the global information derived from the preview. This is in accordance with the findings of Saida and Ikeda (1979) who observed decreased task performance in a moving window paradigm with a small window size. However, with only local information available, the humans in the study of Castelhano and Henderson (2007) managed to perform the search task, and the performance was not as deteriorated as one would have expected considering the findings of Saida and Ikeda (1979). The results of Brockmole et al. (2006) indicate as well that local information significantly contributes to the whole process of object search.

The presented research thus shows that global information guides eye movements and is an important contributor to object search and target identification, and it also guides eye movements. Still, the effect of local information is not negligible and needs to be taken into consideration. Most likely, local information refines the perceived gist of a scene and serves fine-tuning purposes, both for perception and eye movements. Unfortunately, only very limited information is available as to how global and local information exactly play together.

5. Semantic Influences

At first glance it might seem that semantic influences should have already been covered and included in the top-down aspects of scene perception and eye movements, but actually this is not the case. Consider a visitor in a museum looking at paintings. Although possible, the visitor most likely does not have a particular task to perform and watches the paintings for pure pleasure. Still, different objects in a scene play different roles. Some objects might be conveying a large amount of information with respect to the meaning of the scene whereas others just complement the setting and only serve as artistic supplements. Another category of images, not necessarily found in fine arts, contains objects which violate the scene contexts and are perceived as not belonging in a particular image. In order to gain further insights into how the human brain processes visual information, some researchers have looked at how informative or semantically inconsistent objects influence eye movement behavior.

According to Rayner and Pollatsek (1992), the eyes move to informative objects in a scene and remain there until the object is satisfactorily identified. Unfortunately, there is no authoritative measure for informativeness. Mackworth and Morandi (1967), who were trying to establish a relationship between informative scene regions and eye movement behavior, had observers rating scene regions based on their informativeness. Informativeness of a region was defined as the likelihood of recognizing the region at another occasion. Later on, eye movements were recorded for different observers and compared to the initial ratings. The results showed that regions that had been rated informative had higher fixation density than uninformative regions. Some of these did not get any fixations. From there it was concluded that peripheral vision allowed distinguishing informative from non-informative scene regions (Henderson et al., 1999) because the information about semantic consistency or inconsistency must have been available prior to object fixation. In a similar experiment, Antes (1974) defined informativeness as the amount to which a region contributed to the meaning of an image (Henderson et al., 1999).

Summarizing this line of research, Rayner and Pollatsek (1992) assume that important or unusual objects are usually fixated early over the course of viewing, which indicates that significant semantic information must be captured prior to object fixation. In contrast, Henderson et al. (1999) have shown that the eyes are not initially drawn to objects which are inconsistent with scene gist, contradicting the conclusion of Rayner and Pollatsek (1992). However, Henderson et al. (1999) observed that fixation density is higher on inconsistent objects, and that eyes tend to return to inconsistent objects. The results lead them to the conclusion that eye movement behavior changes over the course of scene viewing, as "the eyes are initially driven by visual factors and global scene semantics, with cognitive and semantic aspects of local scene regions playing an increasingly important role as scene exploration unfolds." (Henderson et al., 1999, p. 11)

C. COMPUTATIONAL MODELS OF VISUAL ATTENTION

After having described the cognitive aspects of attention and eye movements in various settings, several computational models of eye movement and attention are introduced.

1. A Saliency Based Visual Attention Model

Probably the best-known model of visual attention has been developed by L. Itti et al. (1998) and is mainly based on the earlier conceptual model of Koch and Ullman (1985). The model follows a pure bottom-up approach in which salient locations are computed from a set of visual features. This saliency is used to determine the focus of attention. Salient locations are characterized by their difference from the background. Salient locations stand out and capture human attention. Humans watching an image with a green dot on a red background tend to first focus on the green dot. This green dot captures the attention of the observer. Treisman and Gelade (1980) have shown that objects which are unique in one feature dimension (e.g., color) among distractors immediately capture the attention of observers as well . This phenomenon has been called pop-out because the time which elapses before the distinct object is focused on is constant and independent of the

number of distractors. This performance can be achieved because the features that yield the pop-out phenomenon can be extracted pre-attentively and no serial scanning of items or attention allocation to the presented items is necessary. Koch and Ullman (1985) name color, orientation, curvature and stereo as such pre-attentive features. Navalpakkam and Itti (2005) also list size, closure, intensity, flicker, and direction of motion. For the first implementation of the saliency-based visual attention model color, intensity and orientation have been picked as the visual features for saliency computation (L. Itti et al., 1998). These are the most tangible ones from a computational perspective, and they are appropriate for evaluating static scenes. In later incarnations of the model, more features have been integrated such as flicker and motion (L. Itti, Dhavale, & Pighin, 2003). These are apparently of relevance for image sequences or videos only.

The saliency-based model of visual attention is heavily biologically inspired and tries to closely mimic the function of the primate visual system. Based on the idea of a two-staged human visual system consisting of a pre-attentive stage and an attentive stage, the model performs computations similar to the mechanism of the pre-attentive stage and generates the focus of attention as output. The following description of the model is based on L. Itti et al. (1998) unless otherwise noted. The model essentially consists of five steps:

1. The visual features are extracted across the complete scene, one channel for each feature (intensity, color and orientation).

2. The center surround differences are computed for each channel at 6 different spatial scales.

3. For each channel the 6 scales are combined and the results are normalized.

4. These normalized maps are combined into the saliency map.

5. The focus of attention is determined based on a winner-take-all strategy.

These five steps will now be looked at more closely.

a. Feature Extraction

First of all, the basic visual features are extracted. The features are intensity or luminance, color, and orientation. The intensity of one image-pixel is computed by adding up all RGB color values and dividing by three as defined by the following formula.

$$I = (r + g + b)/3 \tag{4}$$

where r, g, and b are the color values in RGB format for red , green, and blue respectively. Then four color values, red, green, blue and yellow are computed for every pixel in the following way:

$$R = r - \frac{g + b}{2} \tag{5}$$

$$G = g - \frac{r + b}{2} \tag{6}$$

$$B = b - \frac{r + g}{2} \tag{7}$$

$$Y = \frac{r + g}{2} - \frac{|r - g|}{2} + b \tag{8}$$

The reason for computing these four color values will become clear when the next stage, the center-surround stage of the model is discussed. From the original image 5 additional images or maps (1 intensity, 4 color) with the same size as the original image have now been created. For each of these maps 9 Gaussian pyramids with scales from 0 to 8 are computed where scale 0 is the initial map. Each subsequent scale or layer of the pyramid is derived from the previous scale by smoothing with a Gaussian filter and subsequent down-sampling. No information is provided for the size and the parameters of the filter. Down-sampling is performed by leaving out every other row and column. This means the map on the subsequent layer has half the width and half the height of the previous layer. This mechanism of smoothing and down-sampling serves two purposes. First, it removes noise from the image and second, the different layers can be considered to encode different kind of information. The finer scales still contain all visual details whereas

the coarser scales lose detailed information and thus capture more background information. This distinction will be employed in the next step of the model.

For the orientation feature, Gábor pyramids are calculated from intensity data on 9 scales and 4 orientations (0°, 45°, 90°, and 135°). The map sizes on each of the pyramid layers match to the sizes of the color and intensity pyramids. The computation of the orientation pyramids is performed with the method introduced by Greenspan et al. (1994).

b. Center-surround Mechanisms

In the next stage center-surround computations are performed which are supposed to resemble center-surround mechanisms that can be found at several levels of the human visual system. The center-surround computations are performed for every feature channel on six spatial scales. The center- surround comparison is achieved by point-wise subtracting maps of different spatial scales from the feature pyramids. As mentioned earlier, the finer scales represent detailed information and the coarser scales represent the characteristics of the background. According to L. Itti et al. (1998) the differences of two maps at different spatial scales sufficiently represent the center-surround mechanisms of neurons along the human visual pathway. Since coarser maps contain fewer pixels, they first are converted to the finer scale through interpolation. Each pixel on the coarser scale represents the surround at this location since it contains information derived from the neighboring pixels, at first through the filtering and then through the interpolation.

In the human visual pathway, the center-surround mechanism of color-sensitive neurons operates on color opponency. These neurons are excited when the center of the receptive field senses one color and the surround senses the opponent color. They are inhibited in the reverse case. Four types of such color opponency neurons exist:

- Sensitive to red in the center and green in the surround.
- Sensitive to green in the center and red in the surround.
- Sensitive to blue in the center and yellow in the surround.

- Sensitive to yellow in the center and blue in the surround.

The model subsumes the first two types into the red/green color opponency and the latter two into blue/yellow color opponency. Center-surround maps are thus computed for intensity I, red/green opponency RG, blue/yellow opponency BY and for the four orientations O. The center-surround computation is performed by point-wise subtraction of coarse scales from fine scales denoted $\ominus$. Each pixel at the fine scale assumes the role of one center and the corresponding pixel on the coarse scale assumes the role of the matching surround. Since maps of different scales differ in the number of pixels, the coarse maps need to be filled with pixels. The values for these pixels are derived by interpolation.

$$\begin{aligned} I(f,c) &= |I(f) \ominus I(c)| \\ RG(f,c) &= |(R(f) - G(f)) \ominus (G(c) - R(c))| \\ BY(f,c) &= |(B(f) - Y(f)) \ominus (Y(c) - B(c))| \\ O(f,c,\theta) &= |O(f,\theta) \ominus O(c,\theta)| \end{aligned} \tag{9}$$

Where f refers to the fine scale and $c = f + \delta$ refers to the coarse scale, and $f \in \{2,3,4\}$, $\delta \in \{3,4\}$; $\theta \in \{0°, 45°, 90°, 135°\}$.

This yields 42 feature maps: 6 intensity maps, 12 color maps and 24 orientation maps. One-third of the maps are at scale 2, 3 and 4 respectively. It is important to note that the intensity and color center surround computation is performed by subtracting and returning the absolute value of the result. This means that bright pixels in front of a dark background are treated the same way as dark pixel are on a bright background. The same is true for red/green and green/red opponency as well as for blue/yellow and yellow/blue opponency. This not only violates the resemblance to human neurophysiology but can also result in strange model behavior as was demonstrated by Frintrop (2006) (see II.C.3 on page 47 for more details). Another peculiarity can be seen in the computation of center-surround maps for orientations. This is biologically implausible and also does not make sense from a computational standpoint. The definition of an orientation does not make any sense without considering the surround. Thus, the orientation feature implicitly encodes

center-surround effects, and the additional center-surround computation seems to be redundant. Possibly it has been introduced to be consistent with the computation of intensity and color center-surround maps.

The resulting feature maps are normalized. The normalizing consists of two steps. First, all values are normalized to a pre-specified range [0,M]. Second, each map is multiplied with a factor such that a single local maximum among an otherwise rather uniform surround is promoted strongly, whereas a set of local maxima is suppressed. This becomes plausible considering the following example. One single red dot on a green background should be considered more salient than one red dot among a multitude of red dots on a green background.

c. Computing the Conspicuity Maps

In the next step the normalized feature maps of each channel are added pixel by pixel. All maps are down-sampled to scale 4 by removing every other row and column from finer levels until the map size matches the one of scale 4. This is the coarsest scale of the center-surround maps and therefore no up-sampling is necessary. The resulting maps are called conspicuity maps and are nothing else than saliency maps for a particular feature channel.

d. Creating the Saliency Map

The three conspicuity maps are normalized again using the same mechanism employed for the normalization of the center-surround maps. Finally, they are added up with equal weights, and thus the saliency map is derived. An example of a saliency map can be found in Figure 2. Bright spots indicate high saliency.

e. Determining the Focus of Attention

In a winner-takes-all strategy the location with the maximum saliency gets the focus of attention. This location is suppressed for a certain amount of time, which

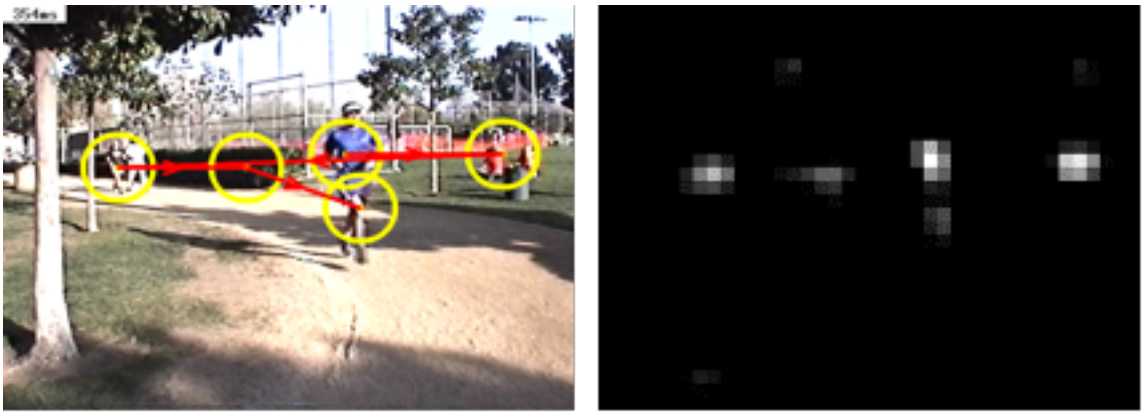

Figure 2: On the left an image with subsequent foci of attention and on the right a saliency map. (From: ILab homepage, `http://ilab.usc.edu/toolkit/screenshots.shtml`, last accessed 04 JUN 2009)

serves as 'inhibition of return' such that the following focus of attention is deployed to the next most salient location. This yields a gaze pattern fixating multiple locations in turn.

f. Discussion

L. Itti and Koch (2000) claim that the model showed great performance in a lot of circumstances. The results are qualitative in most cases, for example, pictures in which the focus of attention is on a plausible location or marks the target that would have been hard to detect otherwise, as can be seen in Figure 3.

In cases in which quantitative results are presented, the performance of the models is judged as good by showing faster response times on search tasks than humans. This, however, cannot be considered as good performance if the model is supposed to resemble human behavior. L. Itti et al. (1998) do not specifically claim that their model actually does resemble human behavior; however, the model is also used to animate head and eye movements for avatars (L. Itti et al., 2003). One would expect that for this task the model should resemble human behavior. Although it does not, it might still be good enough for the mentioned purposes, but it cannot be considered a realistic model of human attention deployment.

Figure 3: Fixation location in a search task. (From: L. Itti & Koch, 2000)

Furthermore, in some initial experiments with the scenes used in the target detection experiment of (C. Darken & Jones, 2007), the model did not allocate attention to any one normal colored target. An example of the model behavior in one of these scenes is shown in Figure 4. The image shows the attended locations after 6.7 seconds. It took human subjects 3.6 to 4.7 seconds to detect the target. These tests used the standard settings of the model and did not involve tweaking the weightings of the conspicuity maps or tuning parameters. One would not expect optimal performance in this case, especially for scenes in which humans showed bad performance as well, but for scenes with detection rates close to 100% one would expect the model to allocate the focus of attention at least a few times to the target. This does not occur as illustrated in Figure 4.

Further critique of the model comes from Henderson et al. (2007). They recorded eye movements of observers viewing images in an object search task. The fixations of the observers were compared to each other and also to the model using two different measures of similarity as described earlier. The average similarity index between human subjects was significantly higher than the average similarity index between the model and the human observers. According to Henderson et al. (2007) the model does not perform

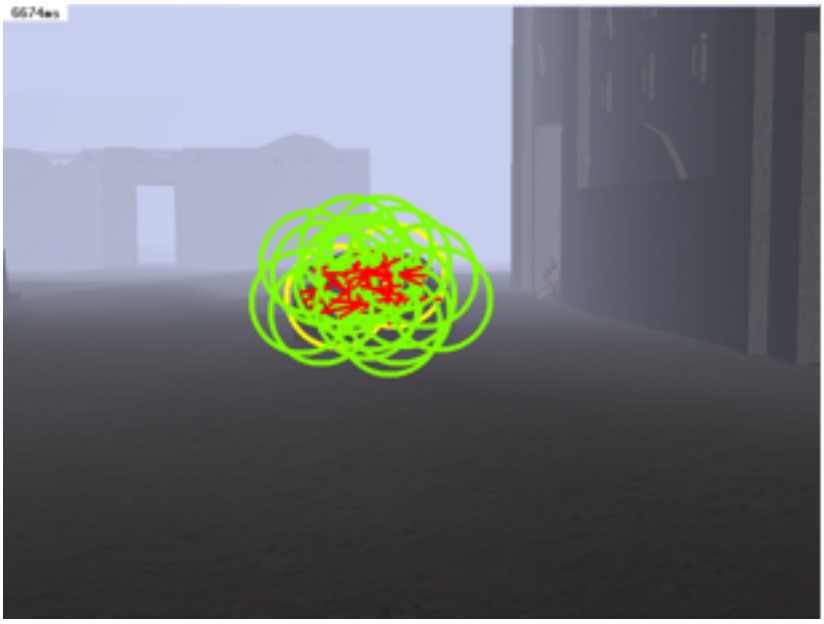

Figure 4: Fixation locations in a search task. The target in the doorway to the right does not receive a single fixation.

well in a search task in real world scenes. This is not too surprising since the humans had a specific task to perform and the saliency model does not incorporate any task influence.

Overall, this shows that it is not appropriate to use the saliency-based visual attention model to predict eye movements of humans who perform a task. Thus, a pure bottom-up model is not sufficient to model human eye movement behavior and it is necessary to take task influences into account. Currently, there are only very few models of visual attention which include top-down aspects. In the next section two such models developed by the research group led by Laurent Itti are described.

2. Task Dependent Extensions to the Saliency Based Visual Attention Model

The original model of saliency-based visual attention of L. Itti et al. (1998) does not take any task-dependent influences into account but relies only on bottom-up features. As described in previous sections (II.B.3 on page 26 and II.B.5 on page 32), the deployment of attention is heavily influenced by top-down mechanisms. The following two models

extend the saliency-based model in different ways that both represent task knowledge to some degree.

a. Top-down Modulation of Visual Features

For the first of the two models developed by Navalpakkam and Itti (2005) the authors claim that they model the influence of task on attention, but the essential parts of the model concern just top-down modulation of a bottom-up model for visual attention. The main difference between these two mechanisms is that a task-dependent model extracts task-relevant information from the scene and uses that to guide attention. Top-down modulation biases the bottom-up features such that visual features prominent in task related objects are strongly promoted whereas others do not receive such promotion. This is especially suitable for searching objects with known visual attributes, and it seems that the main purpose of the model is object detection and recognition.

For detecting objects the model first learns a representation of the object based on the pyramid features as described in the previous section. From this representation, weights of how the individual features contribute to the construction of the saliency map are computed. In order to detect an object, bottom-up visual attention is used. The saliency map is computed using the weightings derived in the previous step, and shifts of attentions are based on this saliency map. At each focus of attention the model computes the features of the object and tries to match them to stored object representations.

To learn the object representation, multiple feature vectors are computed for one object. These multiple feature vectors are derived from several images containing the object in front of different backgrounds and from several viewing angles. In each training image the object is masked and feature vectors are only computed for this masked area, i.e, the area where the object is located. Not only one feature vector is computed, but the object area is divided into a 3 x 3 grid. For each cell of the grid one feature vector at the center of the cell is computed. The derived feature vector for one cell is called a "view." Computation of the feature vector is based on the feature pyramids of the bottom-up attention model described in the previous section, which are 7 pyramids, 1 intensity pyramid, 2

color pyramids (red/green and blue/yellow opponency) and 4 orientation pyramids. From every pyramid, 6 center surround maps at different spatial scales are computed. In total this yields 42 different center-surround maps. For each map the mean and standard deviation are computed. This yields a 42-dimensional feature vector of means and standard deviations for each view. The feature vectors of the 9 views are combined into a unitary object representation. The combined representation is again a 42-dimensional vector of means and standard deviations derived from the input feature vectors and based on the assumption that the features are normally distributed and independent of each other.

At first, views are combined into object instances, where several instances of one object are combined into one representation of that object, several objects are combined into object classes, and so forth. This results in a tree-structured hierarchy with views as leaves and a non-specified, abstract super-object class at the root. The representations derived from the combination step (vectors of means and standard deviations) are stored as the nodes of the tree-structured hierarchy of object classes.

Deployment of attention for detecting an object is performed by using the saliency maps of the bottom-up model. The difference in this model is based on the weightings of the maps, which are added up to form a new map all the way up to the saliency map. This starts from the computation of the conspicuity maps. When the conspicuity maps are derived from the seven feature channels each channel gets assigned a weight or relevancy. This relevancy is computed using the mean and sigma of the respective channel from the object representation corresponding to the object which is to be detected:

$$R(f) = \frac{\mu(f)}{1 + \sigma(f)} \tag{10}$$

The relevancy for each conspicuity map is proportional to the maximum of relevancy for the contributing feature channel. This biases the model to assign high saliency to visual features which are characteristic for the object that needs to be detected.

At the salient location, object recognition is performed. This object recognition computes the object representation of the attended location, i.e., the means and stan-

dard deviations of the feature vectors. Starting at the top of the tree-structured hierarchy the object representation of the attended location is compared to all stored representations at the top level of the hierarchy. If a good match is found, the objects in that branch on the next level of the hierarchy are examined. This is done until the lowest level of the hierarchy is reached. This means that there is an exact match with an object instance. If the evaluation on a certain level only produces worse matches than for the node on the previous level, this node will be the result of the object recognition (Navalpakkam & Itti, 2005).

This model clearly focuses on object detection and recognition. This becomes very apparent when looking at the report of the test results. The main measure of performance is the speed of detection compared to the pure bottom-up model. This speed does not necessarily reflect the behavior of a human. No comparison to human performance on similar scenes is provided, and thus it is impossible to judge how suitable this approach is for a human behavior model. Additionally, the approach does not use actual task influence. It is better described as a top-down modulation of a bottom-up model of visual attention. Although this model seems to be a huge improvement to the previously described pure bottom-up model, it only seems to be valuable in situations that ask for rapid object detection and recognition.

b. Eye Position Based Learning

The second model of visual attention which takes task influence into account also makes use of low-level features for predicting fixation sites (Peters & Itti, 2007). It does not use a tweaking of the saliency model however, but instead uses an interesting learning approach. The idea of the model is to learn a relationship between low-level scene features and eye positions of humans. The model consists of a training phase and of an application phase, in which the trained model is used. For the training phase human subjects are required to perform a specific task, in this case playing video games (Peters & Itti, 2007).

The eye positions of the participants are recorded and feature vectors for all scenes are computed. The eye positions are translated into a so-called gaze density map.

This map consists of 300 elements which correspond to the 300 cells of a 20×15 grid the scene image is subdivided into. The values in the gaze density map are the numbers of eye positions in the corresponding cells of the grid on the image. Two possibilities of extracting feature vectors to test the model are described. The first possibility is to use a 448-dimensional feature vector based on the feature pyramids of the saliency-based model. The other possible feature vector contains 384 dimensions which are derived using a fast Fourier transformation of the image. The description in the following paragraphs will focus on the first feature vector.

The 448 dimensions are derived by using the luminance pyramid, the two color opponency pyramids, and the four orientation pyramids. From each pyramid 2 layers (scale 2 and scale 5) are extracted. These 2 layers are subdivided in 16 cells of a 4×4 grid. For all cells, the means and standard deviations are computed. These means and standard deviations are the elements of the feature vector. For 2 scales for each of 7 pyramids with 16 cells and 2 values per cell this yields $2 \cdot 7 \cdot 16 \cdot 2 = 448$ elements per feature vector. The learning is performed employing a linear least-squares best fit between gaze density vectors and feature vectors. This requires a multivariate regression to be performed with a $T \times N$ response matrix of gaze density maps and a $T \times M$ input matrix of feature vectors where T is the size of the training set, N is the size of the gaze density map or gaze density vector ($20 \cdot 15 = 300$ elements), and M is the size of the feature vector (448 for the pyramid features). In the learning phase the regression matrix is computed based on the recorded eye positions and the feature vector. In the application phase a gaze density map for one scene image is computed by multiplying the regression matrix with the feature vector of the scene image.

To test the model, the derived gaze density map is compared to actual eye positions of a test set which was not contained in the learning set (Peters & Itti, 2007). Comparing the model results to a purely saliency-based model it becomes apparent that the predicted gaze density is a better indicator for actual eye position. A combination of the saliency map and the predicted gaze density map, derived from point-wise multiplication of

the two maps, performed even better. Peters and Itti mention that there are other potential learning mechanisms, and that a better mechanism of combining saliency and gaze density would be desirable. This again points out that there is a general lack of understanding how the combination of these two mechanisms has to be performed and how exactly they interact.

Peters and Itti (2007) also point out, that one weakness of the model is that it has to rely on the training set. A particular training set does not necessarily allow generalizing the model to any situation. According to Peters and Itti (2007) it is unclear yet how large a training set has to be to allow generalization to any situation. Furthermore, the model has very coarse gaze density maps. A subdivision of the images into 20×15 grids can easily mean that each cell subtends more than one degree of visual angle. Thus, any fixation location would be constrained by this limited resolution. Of course the resolution of the gaze density map can easily be increased for the cost of higher computational demands. More importantly, the model captures task influences indirectly only by establishing a relationship between low level visual features and eye movements. This assumes that the low level features used are representative of the information humans capture from a scene to perform their task. This needs to be evaluated before being able to judge about the quality of the model. Also, the model assumes that a rather simple learning mechanism is able to represent the connection between eye movements and task knowledge. Although the authors mention that there are better learning mechanisms, it cannot be assumed that any one of them automatically captures the actual relationship of eye movements and task information. A thorough investigation of this relationship needs to be conducted in order to allow for the decision on an appropriate learning mechanism.

Although not explicitly stated, it seems safe to assume that eye movements could be performed with an approach similar to the one used for the saliency-based approach. This means the winner-take-all approach would be employed to allocate the gaze to the location with the highest predicted gaze density. However, this implies that a new eye position would be generated for each new frame. These eye positions might jump

around on the scene any distance, which is in contrast to psychophysical evidence (Rayner & Pollatsek, 1992). Additionally, the model integrates over the behavior of a large set of people. Although this is generally a good idea, it levels out any cognitive factors or emotional factors. For a technical application this might be the desired effect, but for a human behavior model these influential factors need to be captured, and the model should be able to react to these inputs given they can be provided. Theoretically, one could try to stimulate subjects accordingly and establish different regression matrices for different cognitive and emotional states. However, this does not seem to be easily feasible. To summarize, it can be stated that this model nicely establishes a relationship between task and eye movement which seems to be very suitable for technical applications. Using it for human behavior modeling should only be done very carefully, if at all, but it might inspire future implementations of human behavior models.

3. VOCUS

Although the Visual Attention System for Object Detection and Goal-Directed Search (VOCUS) developed by Frintrop (2006) is heavily based on the visual saliency model of L. Itti et al. (1998) it is worth describing. Frintrop identified some weaknesses of the Itti model which her model alleviates. She also introduces some simplifications on the implementation side which leads to somewhat slower processing times but reduces complexity. In the following, the important changes and the corresponding motivations and implications are described.

a. Feature Extraction

The features used for computing the saliency map are intensity, color, and orientation. These are the same features as the ones used by the Itti model, but they are computed differently. The intensity values are computed by converting the input image to grayscale and using the grayscale values for luminance. Unfortunately, it is not mentioned how the conversion to grayscales is performed. However, this cannot be considered to be a major difference to the luminance computation of the Itti model. From the resulting

grayscale image a Gaussian pyramid with 5 layers $s_0 ... s_4$ is computed where s_0 is the original grayscales image, and s_{i+1} is derived from S_i by filtering it with a 3×3 Gaussian kernel and down-sampling the filtered image, i.e., dropping every other row and column.

A more significant difference can be found in the computation of color values. Frintrop converts the input image to the LAB color space. The big advantage of this color space is that the color channels encode color opponency for green versus red in the A-channel and for blue versus yellow in the B-channel meaning that smaller values indicate "greener" or "bluer" hues and larger values indicate "redder" or "yellower" hues respectively. A middle value indicates colorlessness, i.e black or white depending on the luminance. This representation of color lends itself quite naturally to the color opponency mechanism in retinal ganglion cells and other neurons on the visual pathway. From the image in LAB color format a Gaussian pyramid with 5 layers is computed similarly to the one for intensity. This pyramid is used to generate 4 pyramids for the 4 colors red, green, blue and yellow respectively. Each pixel in the maps of the originally computed color pyramid is a three-dimensional vector containing a value for each luminance, A-value and B-value. The pixels in the pyramids of the individual colors are scalars representing the distance to the prototype of the respective color. Both the red/green and the blue/yellow axis have a range of [0 ... 255] where 0 means green-most or blue-most and 255 means red-most or yellow-most respectively. A pure color of red, green, blue, or yellow means that the value in one color dimension assumes one of the extreme values and in the other color dimension it assumes the mean between the two extremes. Ignoring intensity, the red prototype for example would be the tuple (255, 127). Thus, the pixel values for a particular pixel on a particular layer of the red pyramid would be derived by computing the Euclidean distance between the according pixel on the according layer of the LAB pyramid and the value of the red prototype. The computation operates on 2 dimensions, A and B only, thus ignoring the intensity value.

$$
\begin{aligned}
P_{C,s} &= d\left(P_{LAB,s}, c\right) \\
&= \left\| P_{LAB,s}\left(x, y\right) - c \right\| \\
&= \left\| \left(p_{A,s}, p_{B,s}\right) - \left(c_A, c_B\right) \right\| \\
&= \sqrt{\left(p_{A,s} - c_A\right)^2 + \left(p_{B,s} - c_B\right)^2}
\end{aligned}
\tag{11}
$$

$P_{C,s}$ is one pixel on layer s of the pyramid for the color C, $C \in \{red, green, blue, yellow\}$, $P_{LAB,s}$ is the according pixel on layer s of the LAB color pyramid, and c is the value for one of the prototypes red (255,127), green (0,127), blue (127,0) or yellow (127,255) according to C; $p_{A,s}$ and $p_{B,s}$ are the values of the A and B channels on layer s of the pyramid. The result of these computations are 4 pyramids, one for each of the colors with 5 layers per pyramid, i.e. 20 maps.

Like in the Itti model. In VOCUS the method of Greenspan et al. (1994) is used for the computation of oriented pyramids. VOCUS uses one orientation pyramid for each of the orientations 0°, 45°, 90°and 135°.

b. Center-surround Mechanisms

In contrast to the Itti model, the center-surround computation is not done across pyramid scales but on individual scales, and for scales $s_2 ... s_4$ only. Leaving out the two largest scales is due to the desire for noise reduction. The center surround computation is performed by first computing the average of all pixels within a given radius from the center. Two radii, either three pixels or seven pixels, are used for determining the range of the surround. Then, two center-surround maps are computed for each combination of the three scales and the two surround sizes, an on-center center-surround map and an off-center center-surround map. The on-center map yields a high response for a bright center on a dark background, and the off-center map yields a high response for a dark center on a bright background. This can be achieved by calculating $center - surround$ in the first case and $surround - center$ in the second case with every pixel of each scale being in

the center once. The exact algorithm is unclear. A similar approach is used for the center-surround maps of the color channel. For each color pyramid the on-center computation is used to derive the difference for the respective color to its background. The background computation and scales are the same as for the intensity channel.

Frintrop (2006) uses this method of computing the center-surround for two reasons. First, the method of the Itti model does not allow feature pop-out in some cases. A bright spot on a grey background and a dark spot on a grey background are effectively indistinguishable in the Itti model because of the absolute value $|\mathtt{center} - \mathtt{surround}|$ in the center-surround computation. VOCUS makes a distinction between on-center and off-center responses, which will achieve a pop-out for one white spot among black spots on a grey background. The second reason is that a top-down biasing for dark-on-bright would be the same as for bright-on-dark, which apparently is not sufficient.

Interestingly, no explicit center-surround computation is performed for the orientation pyramids. Frintrop (2006) claims that an implicit center-surround computation is already performed during feature extraction.

The center-surround maps for layers $s_2 ... s_4$ are summed up for both the on- and off-intensity maps and for all 4 color maps, yielding 2 intensity maps and 4 color maps. Similarly, the orientation maps on layer $s_2 ... s_4$ are summed up for all 4 orientations for a total of 4 orientation feature maps. After this step there are 10 feature maps that will be combined into conspicuity maps in the next step.

c. Computing the Conspicuity Maps

One conspicuity map is computed for each feature channel. Before this can be done the maps within one feature channel have to be normalized to make sure the saliencies in the different maps contribute adequately to the conspicuity maps. Frintrop points out that the normalization method proposed by L. Itti et al. (1998) strongly promotes single peaks in a map and suppresses two local maxima, even if they stand out significantly from the background. To alleviate this, Frintrop suggests a different way of normalization she calls uniqueness weight (Frintrop, 2006). This weight is derived by dividing a particular

map by the square root of the number of local maxima above a certain threshold. Thus, maps with a higher number of salient regions contribute less to the conspicuity of the feature. In the example of the one bright dot and several dark dots on a grey background, the values of the on-intensity map will have larger values than the off-intensity map. The on-intensity map has only one salient location, and thus it will be normalized with factor 1, whereas the off-intensity map has 7 salient locations and the map will be normalized with factor $1/\sqrt{7}$. This will result in the bright spot being the most conspicuous location for the intensity channel. This step is identical for all three feature channels and yields three conspicuity maps which are fused into the final salience map.

d. Computing the Saliency Map

In order to derive the final saliency map all three conspicuity maps have to be combined. This is achieved by adding up all three conspicuity maps, but not before the conspicuity maps are normalized. Since the conspicuity maps are derived from a different number of feature maps, their values will be in different ranges. Thus, the normalization needs to make sure that no feature channel is inappropriately promoted a priori over the others. In the Itti model the conspicuity maps are normalized to a fixed range. Frintrop (2006), however, points out that one channel might stand out in comparison to the other channels due to the nature of the scene. This difference should not be leveled out in the normalizing step. Thus, instead of using a fixed maximum for all conspicuity maps, in VOCUS, for each feature channel (intensity I, orientation O, color C), the maximum of all feature maps ($\hat{m}_I, \hat{m}_O, \hat{m}_C$) is computed and the respective conspicuity map is normalized in the range $[0 \cdots \hat{m}_f]$ where $f \in \{I, O, C\}$. The final saliency map is computed by applying the aforementioned weighting function to each of the normalized conspicuity maps and adding them up.

$$S = W(I) + W(O) + W(C) \tag{12}$$

e. Discussion

Several very plausible critiques of the Itti model are presented by Frintrop (2006). Furthermore, suggestions to alleviate these issues are presented. The solutions are reasonable and the provided examples indicate that they fulfill the intended purposes. Unfortunately, for VOCUS there is no source code freely available. Thus, it is not possible to directly test the model on the same search scenes the model of L. Itti et al. (1998) was tested against. It is therefore difficult to judge whether the improvements of VOCUS compared to the Itti model will result in better performance when used on the search scenes of C. Darken and Jones (2007).

4. Contextual Guidance Model

The contextual guidance model of Torralba et al. (2006) strives to combine bottom-up and top-down aspects for the deployment of visual attention in object search. The top-down information they include is based on global information as described earlier. For the representation of global information a specific mechanism for the computation of global features is introduced. These global features are used to establish a relationship between object locations and global information. This allows the determination of locations which are likely to contain an object based on the global information, or more precisely global features, but local features also contribute to the likelihood for object presence at a location. The local features are essentially the same as the visual scene features extracted by the visual attention model of Itti (II.C.1 on page 33) or by VOCUS (II.C.3 on page 47). The combination of bottom-up and top-down mechanisms is based on a probabilistic framework which will be described next.

The actual task in visual search requires an observer to determine whether the specified target is present or absent and to indicate the target location. The contextual guidance model assumes that the observer will scan the locations based on their probability of containing the target. Locations with a higher probability will be scanned first. Thus, $p(O, X|I)$ has to be computed for every location X, with O indicating target presence

(1=target present, 0=target absent) and I being the image features. In order to derive the fixation locations, the features need to be extracted attention free or pre-attentively from the scene. The used information is local features derived from a bottom-up mechanism and global, contextual features, which can be extracted pre-attentively as well. The local features are denoted L, and the global features are denoted G and according to the theory $p(O, X|I) = p(O, X|L, G)$. Using the laws of conditional probabilities the probability can be broken up into four terms:

$$p(O = 1, X|L, G) = \frac{1}{P(L|G)} P(L|O = 1, X, G) P(X|O = 1, G) P(O = 1|G) \quad (13)$$

These terms can be interpreted as mechanisms contributing to the allocation of attention. $P(L|G)$ does not contain any target relationship and is a pure bottom-up factor. It represents the likelihood of finding specific local features in a given image. Thus, $1/P(L|G)$ fits to the definition of saliency, according to Torralba et al. (2006). The second term $P(L|O = 1, X, G)$, incorporates the knowledge of the visual features of the target and is thus a top-down factor. It is the only factor containing any information about target appearance. $P(X|O = 1, G)$ indicates likely target locations based on the context. It is a factor that is based on experience and expectancies of an observer. Finally, $P(O = 1|G)$ tells how likely it is that a given scene contains the target at all. This information is excluded from further consideration since its influence is considered constant for the intended purposes of the model. Also, the influence of target appearance $P(L|O = 1, X, G)$ is excluded. The reason for that is based on the assumption that this factor requires deployment of attention to actually integrate the features into a unitary percept, but the purpose of this model includes determining the initial fixation or initial spot of attention. Thus, any computation which would require attention is not supposed to contribute to the model output. Although this argumentation seems logical, it fails to explain why it is admissible. The initial goal of the model was to mark locations based on their likelihood of containing an object. Removing the first term is admissible given the assumed constancy across the application area. Removing the only term representing target appearance from the model violates the initial

assumptions of the model. It furthermore neglects the possibility of extrafoveal information capture. During the course of scene viewing, information about target appearance is perceived extrafoveally (Henderson et al., 1997), and this information possibly influences eye movements. Therefore, it is questionable whether the output of the model still reflects actual human behavior and human performance. Still, the approach of the model is interesting enough to further examine more details. Without the terms that are excluded the model reduces to the following simplified version:

$$S(X) = \frac{1}{P(L|G)} P(X|O = 1, G) \tag{14}$$

$S(X)$ is called contextually modulated saliency. Apparently, the term probability is avoided since two of the originally derived terms have been left out, and thus the equation does not reflect the probability of target likelihood at a particular location anymore.

Now, the two terms have to be computed. For $P(L|G)$, this is done by first computing local image features and then estimating $P(L|G)$. The local features are computed for the three RGB channels of a color image. Each channel is filtered with steerable pyramids (Simoncelli & Freeman, 1995) for 6 orientations at 4 scales. Basically, this is an edge detection performed on each color channel at 4 different levels of spatial resolution. This computation yields a 72- dimensional feature vector for every location of the image. These vectors are fitted to a power exponential distribution using maximum likelihood. The derived distribution is considered to represent the conditional probability $P(L|G)$. The second term contributing to the contextually modulated saliency represents the relationship between target location and global image features. This relationship can be learned from the global image features of images in a training set with known target locations. The computation of global image features is not done on an image pixel basis but on a coarser level that combines information from multiple locations based on a 4×4 grid superimposed on the image. The features at every location are computed by passing the intensity of the image through a set of steerable pyramids with 6 locations on 4 spatial scales. This yields a 24-dimensional feature vector. For every cell of the 4×4 grid, the average of the feature

vectors contained in that cell is computed. Thus, the global feature vector of an image contains $4 \cdot 4 \cdot 24 = 384$ elements. This vector is reduced to 16 dimensions by performing a principal components analysis on a set of images. The 16-dimensional feature vectors are then computed for a set of prototypical training images with known target locations. The relationship between the features and the target locations is learned. The result of the learning mechanism represents $P(X|O = 1, G)$ and can thus be used to compute the contextual modulated saliency $S(X)$.

The contextual guidance model has been evaluated by comparing its output with the output of a pure saliency-based model and with the results of eye-tracking experiments with human subjects on the same scenes (Torralba et al., 2006). The evaluation showed that the contextual guidance model performs significantly better than the pure saliency-based model. This indicates that the influence of contextual information provides knowledge which increases model performance. However, this comparison has only been done for the first five fixations of each participant. The comparison also looks at how many of those first five fixation land on the 20% of the image at which the map has its highest values. Although this number was statistically significant higher than 20% (i.e., clearly better than chance) it is unclear what the numbers actually mean, how good the model really performs, and how this performance compares to other assessments of eye fixation predictions. In addition, comparing the fixation positions of one individual with the fixation positions of all other individuals and comparing the fixations of all subjects to the fixations predicted by the model, one can see that the model did perform significantly worse. The fixation location of one individual is a better predictor of fixation locations for other subjects than the prediction of the contextual guidance model.

Furthermore, the model still relies on the notion of a saliency map. Also, the global information used for eye fixation prediction needs to be extensively learned and it does not capture semantically relevant content from the scene. Together with the fact, that the eye fixations of other participants are better predictors of the eye fixation of one particular

participant than the model is, this means that more information needs to be taken into account by visual attention models in order to increase their performance.

D. SUMMARY

In this chapter, human aspects as well as computational aspects of visual attention and visual attention modeling have been described. The research examining the human aspects of eye movements and visual attention covered the areas of physiology, neurophysiology, psychology, and psychophysics. Of special interest related to visual attention modeling are the research areas of visual search in abstract search arrays, eye movements in reading, and eye movements in scene perception. This research provided a lot of insights and information for the creation of the currently available visual attention models.

Up until today, the computational models of visual attention and eye movements consider mostly bottom-up information. Top-down contributions are not covered very extensively, although a lot of currently ongoing research is looking at incorporating them into the models. This research mostly focuses on the top-down modulation of the bottom-up saliency maps with the exception of the contextual guidance model, which incorporates global scene information as well.

So far, not a lot of research has been conducted as to how semantically relevant locations influence eye movements. In addition, there is not any visual attention or eye movement model incorporating this type of information. In the next chapter, an experiment examining the influences of meaningful scene locations on eye movements are examined. The chapter afterwards, describes how such information can extracted from a simulation environment and presents the creation of a relevance map. This map represents semantically relevant scene locations and it will be shown, that it predicts eye fixations very well.

III. ASSESSING THE INTERACTION OF BOTTOM-UP AND TOP-DOWN FACTORS ON THE EYE MOVEMENTS IN VISUAL SEARCH FOR A HUMAN TARGET

A. INTRODUCTION

A very important factor for the modeling of eye movements is the interaction between bottom-up and top-down driven influences on eye movements. One aspect of this interaction has been extensively examined within the visual search paradigm by looking at slopes of response times for different sizes of search arrays in pop-out and conjunction search (e.g., Treisman & Gelade, 1980; Wolfe, 1998). Another aspect of the top-down influences has been examined in real world scene perception research by looking at the semantic influence of scene objects on eye movement patterns (Henderson et al., 1999).

However, neither of them sufficiently addresses the question as to how humans move their eyes across real world scenes in the search for a human enemy target to an extent which would allow to directly model the eye movement behavior based on the results of these research areas. The tasks in visual search experiments use only abstract search targets like geometric shapes on uniformly colored backgrounds, and it is not apparent how the results of these experiments could provide answers about the performance in the search for realistic targets in realistic scenes. Real world scene perception research is concerned with realistic scenes, however the interest there is the general eye movement behavior and not the eye movement in search tasks. The goal of the work presented here, on the other hand, is to examine the eye movement behavior in search for a human target in a combat setting. This is different from classical visual search since realistic scenes show higher complexity than classical search arrays and therefore they do not necessarily show the same eye movement patterns. This fact has been recognized in recent years and there is a trend to conduct search experiments in natural scenes (e.g., Einhäuser, Rutishauser, & Koch, 2008; Pomplun, 2006).

Just recently, Pomplun (2006) examined the eye movement behavior in complex stimuli and showed that saccades are directed towards features in a scene which are common to the features of the search target. In this experiment, the search target, which was a rectangular cut-out of the search scene, was displayed to the human participants just before scene onset. In this paradigm, the target features can be encoded prior to the search and subsequently guide the search process in a top-down modulating manner. However, in the experiments of Darken and Jones (C. Darken & Jones, 2007), participants successfully found well-hidden human targets in camouflage uniforms without knowing their appearance beforehand, that is without knowing the size, aspect and visible portion of the hidden targets. Although humans might have general features of humans available for top-down guidance of eye movements, this means that human search strategies can not rely exclusively on the exact visual target features. This idea gets further support by the findings of Henderson et al. (1999) who showed a significant influence of high-level semantic information on eye movements. Similarly, the contributions of informativeness and semantic influence of scene regions on eye movement behavior are of interest for this research. As has been shown by Henderson et al., semantic information of a scene guides the eye movement behavior of observers.

Qualitative analysis of the eye movement recordings conducted by Wainwright (2008) indicates that a substantial number of eye fixations are co-located with possible hiding locations of human targets as opposed to locations attracting the gaze purely based on their visual features. We consider these hiding locations or cover spots to be informative locations with semantic influence on eye movements. Examining under which circumstances these locations attract the eyes of an observer is one goal of the search experiments. The second goal is to examine the interactions of top-down and bottom-up mechanisms for attention allocation. To date, very few experiments have been conducted which examine the interactions of top-down and bottom-up influences on search. Again, classical search paradigms have been used by Theeuwes (2004) and more recently Navalpakkam and Itti (2006). Examining search in real world photographs, Einhäuser, Rutishauser, and Koch

(2008) examined the top-down influences on search and their relationship to sensory-driven signals more closely. Einhäuser, Rutishauser, and Koch found that bottom-up factors are a strong driving factor for eye movements, but these effects are essentially eliminated as soon as observers have to search for a given target. They used grayscale photographs of naturalistic scenes to which a contrast gradient was applied in order to create images with one side being more salient. Also, the natural appearance of the images were varied by applying different amounts of noise to the images. In addition to that, the artificially created search targets, which were added to the scenes, were rather abstract and did not resemble real world objects. One of the targets was a circular Gábor pattern, which is essentially a bullseye target, and the second target was formed through locally increasing the contrast for a small circular area of the image. It is questionable whether the findings of this study will extrapolate to any other naturalistic scene, and therefore we want to further examine the interactions of top-down and bottom-up influences on visual search.

According to Theeuwes (2004) a color singleton, which is a single object being different in color from all other objects, captures the attention of a viewer and interferes with attention allocation to a top-down target. This interference may not take place if the search is serial due to the size of the search array. This view is in opposition to the findings of Bacon and Egeth (1994). They concluded from a very similar experiment that observers can be in two different search modes, namely feature search mode and singleton detection mode. Depending on the mode, the attention capture of the visually salient stimulus can be overridden by task demands. To gain these insights the experimental setup needed to control for participants being in one specific mode. In a laboratory setting this is possible but in a real search task it is impossible to know whether people are in one or the other mode. In addition to that, Theeuwes (2004) argues that there is no such thing as different search modes. Still, there is other evidence that stimulus-driven attentional capture can be overridden by task demands. Einhäuser, Rutishauser, and Koch (2008) showed that the search for a target immediately breaks the stimulus-driven attention allocation.

Due to these contradicting results, there is a need to revisit this issue. This is of even greater importance since the stimuli in the mentioned studies are still too abstract or modified in a non-realistic manner. Also, it is not only of interest whether overriding of attentional capture is possible in general, but if it is possible it is important to know how the task demands compete or interact with the visual features. Specifically, the influence of varying levels of saliency and eccentricity of the targets and distractors on the overt attention allocation, that is the fixations of objects will be examined. The hypothesis is that the first saccade from a pre-cued fixation location will go towards the top-down target. The likelihood of the first saccade being directed to the top-down target will decrease with a higher eccentricity of the target and less visual saliency of the target as compared to the distractor. The expectation is to see an interaction between eccentricity and saliency and the quantitative influences of these factors will be evaluated in order to incorporate them into the computational eye movement model.

The basic presumption is that a clearly visible target close to the fixation location with a salient color distractor at the same eccentricity in the other hemifield will almost always receive the first fixation. The distractor might get the focus of covert attention that will be reflected in the time to fixation of the target, but not in the actual fixation. Increasing the eccentricity and decreasing the saliency or visibility of the target will decrease the proportion of first fixations on the target. The first fixation in this case will go towards the distractor. If at the same time the eccentricity of the distractor increases while its saliency decreases, the first fixation might be on neither the target nor the distractor. We are interested in finding the levels of saliency and eccentricity at which the target does not get the first fixation any more and at which neither target nor distractor get the first fixation any more. This might be a continuous decrease that will be reflected in a reduced likelihood of the first fixation on the target over scenes and subjects, but possibly there will be rather sharp thresholds. It is important to note that hit or miss rates will not be measured in this search experiment. Instead, saccadic selectivity based on saliency and eccentricity of the target and distractor will be measured by looking at time until target fixation and the

number of first target fixations. The response time until participants report target detection will be measured as well.

In addition to that, the influence of semantic guidance factors is to be examined. As could be observed in previous experiments on a qualitative basis, likely hiding locations such as doorframes, window frames, walls, etc. seem to be saccade targets. Apparently, these locations are detected as locations potentially containing the target based on their semantic content for the search task. These locations are considered to be task dependent, actual top-down factors in contrast to top-down modulated bottom- up factors such as red spots in the search for a red target. In this experiment, search arrays will be designed that contain these types of top-down locations in addition to the human target and the distractor object. Again, the goal is to explore where the eyes are guided to, based on the saliency of the target and distractor and the eccentricity of the target, distractor, and semantically important locations.

The intuition is that the target will serve as the most influential factor on saccades. However, this influence will decrease with higher eccentricity from an initially cued fixation location, and then the hiding location will get precedence over the visually salient distractor. That means if the target is at an eccentricity and saliency level such that it does not serve as a saccade target any more, we expect that the hiding location, which possibly could contain a hard to spot target, would get the fixation instead of the distractor, which from a top-down perspective is a useless fixation location. If the hiding locations do not get any fixations or only get fixations very rarely, even in cases in which neither the target nor the distractor are fixated after the first saccade, there would be an indication that hiding locations are not processed as such by the attention allocation mechanisms. This would mean that other factors that coincide with hiding locations, are driving elements which lead the eyes to these spots. These could either be bottom-up influences (for example, edges with sharp contrasts) or top-down modulated stimuli.

In addition to that, we expect that the eccentricities at which targets will immediately receive fixations will decrease as compared to the first part of the experiment. Due

to scenes with higher information content and more clutter, the perceptual span will supposedly be decreased, and thus objects at a higher eccentricity will fall outside of the area being processed by the human attention guidance mechanism. The concept of perceptual span or span of effective vision emanates from eye movement research in reading, and has later been adopted in scene perception research. Perceptual span is the area around fixation from which readers or observers can extract useful information. (Rayner & Pollatsek, 1992)

This study will answer the following questions: how do top-down signals, bottom-up signals and semantic influences guide eye movements and how do they interact in visual search for a human enemy target.

B. METHOD

1. Participants

Nineteen students and faculty of the Naval Postgraduate School in Monterey participated in the experiment after providing informed consent. All participants were members of the U.S. Armed Forces across the four services Army, Marine Corps, Air Force and Navy. The participants were volunteers and did not receive any compensation. All participants were naïve with respect to the hypotheses of the experiment.

2. Stimuli

One hundred six stimuli containing one or zero targets, one or zero distractors, and one or zero locations with semantic influence on the search task are presented to participants. These scenes were designed using a stimulus generation and display application developed at the Naval Postgraduate School based on the Delta3D game engine. The targets used for this experiment were infantry soldiers. The distractor was an unfolded piece of newspaper seemingly attached to a wall behind the target. The location with semantically relevant content was a doorway through which a human could approach or which could be used as a cover spot or hiding location. This will be referred to as the hiding

location from here on. All of these three could have been present or absent but except for the target only trials at least two of the three entities were present.

The size of the target, distractor and hiding location remained constant over the course of the experiment. Within the scenes the target and distractor were varied along two dimensions, that is eccentricity and saliency whereas the hiding location was varied with respect to eccentricity. The eccentricity of these entities could assume four possible levels (0, 1, 2 and 3) with values of 5°, 9°, 13° and 17° of visual angle. These variations occurred along the horizontal axis located at the center of the screen (see Figure 5 for an illustration).

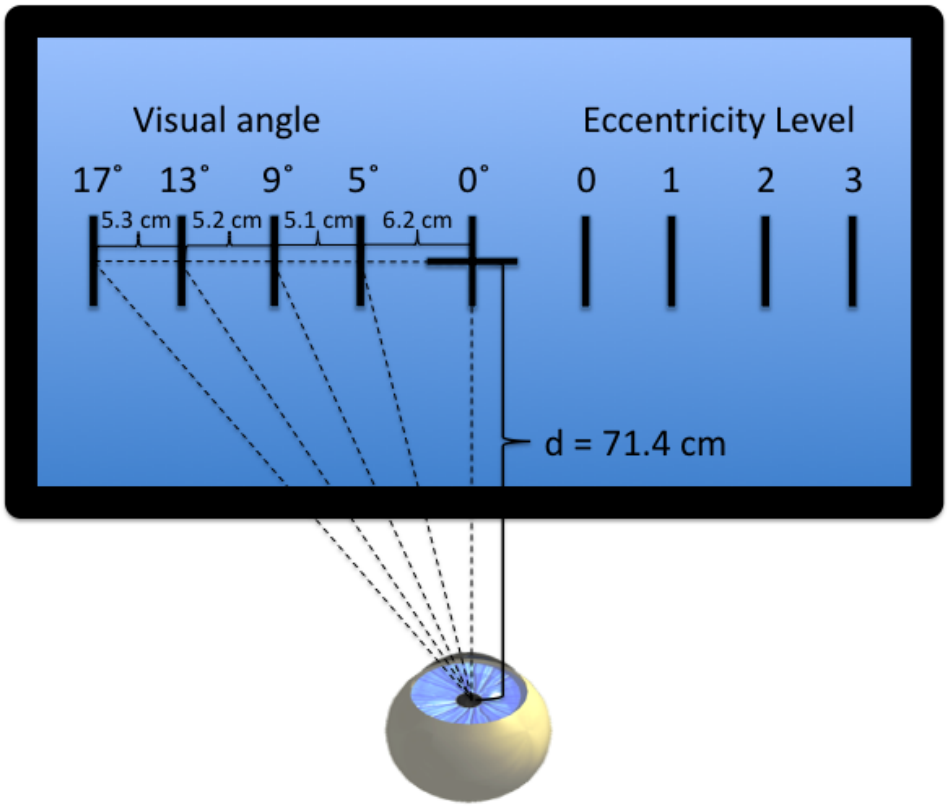

Figure 5: An illustration of the used eccentricity levels. Depicted in terms of degrees of visual angel on the left and in terms of the eccentricity level on the right. The crosshairs in the center indicate the location participants were asked to fixate before stimulus onset.

The eccentricity levels were chosen to cover the whole range of the screen without being too close to the center and without being too close to the edge of the screen. The

objects did not have to be too close to the center such that it was possible to discern a saccade to the target from small movements at the initial fixation location at the screen center. In addition to that, the objects were not placed at the very edge of the screen to avoid any possible influence of the screen frame on the effect of the object for the search task.

The saliency of the target and distractor varied on a scale of the following four possible levels (0, 1, 2, and 3): one low level, one medium low level, one medium high level and one high level. The low saliency level was picked in a way such that the object was almost invisible for the lowest level but could be discriminated from the background when fixated directly. The next higher saliency level, medium low, was set a little higher, so that the object was still hard to see but not as hard as in the lowest setting. At the high saliency level the object was very clearly visible. The medium- high level was set a little lower than the high level. In this setting, the object was still clearly visible but the contrast from the background was reduced as compared to the very high level. The target never assumed the medium-high saliency level and the distractor never assumed the medium low level. All four levels had been visually judged by a human observer. Then, the parameters determining the saliency within the stimulus display software were fixed to ensure equal saliency values for all objects with the same levels. Due to the different nature of the target and distractor the parameters for the same saliency level could have been different for the target and distractor respectively. An example of a stimuli can be seen in Figures 6 and 7.

For both eccentricity and saliency, not all of the four levels were used for all the different entities, and the levels used could have varied from condition to condition. For each condition, the used factor levels will be described in the results section. The used levels were chosen in order to find the right balance between the required number of trials and the required number of factor levels allowing for the assessment of the factor level influences.

Figure 6: A stimulus example containing the target, distractor and hiding location. The target is located at roughly the center of the left hemifield. The displayed contrast is the highest contrast being shown in this condition.

Figure 7: The same stimulus as shown in Figure 6, but the image is modified such that the target becomes clearly visible.

The visual judgement of the contrast level was performed at the same screen, in the same location, and with the same lighting conditions as the actual experiment. The experiment took place in a completely darkened laboratory.

3. Apparatus

The stimuli were presented on a 24 inch TFT monitor set to 60 Hz at a resolution of 1920x1200 pixels measuring 52cm x 32.5cm. The stimulus display software was running on a Dell XPS 720 floorstand PC with a Intel Core 2 Quad processor at 2.4 GHz.

Eye tracking was performed with the Seeing Machines FaceLab4 eye tracker. Eye tracking sampling occurred at 60 Hz and the experiment was only conducted for participants for which the screen calibration resulted in a mean error of 1.0° of visual angle or better.

Participants were placed at a viewing distance of 71 cm resulting in the screen covering a visual angle of 40°. The viewing distance was maintained with a modified chinrest used as a chestrest against which participants leaned during the experiment. The head movements were unrestricted.

4. Design and Procedure

Before taking part in the experiment every participant provided an informed consent. Then, the visual acuity and color vision of participants were tested using a modified Snellen chart and the Ishihara color test respectively. Only participants with an uncorrected vision of 20/30 or better took part in the experiment. With respect to color vision, participants were required to correctly read the charts 1-14 of the Ishihara color test in order to be eligible for participation. In order to increase eye tracking accuracy participants were not allowed to wear glasses or contact lenses during the experiment.

Participants fulfilling the stated criteria proceeded with the experiment. After successful calibration of the eye tracker participants were introduced to the experiment. They were told that their task was to spot enemy targets in camouflage uniform in an urban environment as fast as possible. The participants were asked to find the targets as fast as

possible, fixate on the targets with their eyes, and then press the spacebar to indicate a successful search. If they could not find the target they were asked to say 'next' and the scene would be advanced for them. They were also informed that in addition to the search target, additional objects could appear as would be expected in an urban environment. No further information about the nature of the objects and their meaning for the experiment was provided in order to avoid any biasing or priming with respect to the distractor or the hiding location. Before the start of the experiment the target was introduced to participants in the high contrast setting. Neither the distractor nor the hiding location was shown prior to the experiment.

In order to control for the eccentricity of the entities, a fixation cue was displayed before each scene. This fixation cue, black crosshairs in a white circle on a black background, was located at the center of the screen. Before the experiment the participants were told to look at the crosshairs and do that until the search scene was displayed. Scenes at which the initial fixation was not located within 2° of the scene center were considered errors and excluded from the analysis.

Before the start of the experiment it was made sure that the participants understood the task by asking them questions about the instruction's key points. In addition, two practice trials were performed so that the participants could familiarize themselves with the flow of fixation cues, scenes, and the expected input.

The experiment consisted of the following six conditions: the target-only condition, the target and distractor condition, the target and hiding location condition, the distractor and hiding location condition, and two conditions with the target, distractor and hiding location. The last two differed in the location of the hiding location. In one condition the hiding location appeared in the same hemifield as the target, and in the other condition the hiding location appeared in the same hemifield as the distractor. The target and distractor always appeared in different hemifields. In the other four conditions, which contained only two of the three entities, the two entities always appeared in different hemifields (one on

the left, the other one on the right). In each condition the location of an entity in one of the two hemifields was balanced across trials.

All 106 stimuli were presented in one session without any interruption. The six conditions were presented in two blocks, where the first block consisted of the stimuli without hiding location, and the second block contained all conditions with hiding locations. Within these two blocks, the scene presentation was randomized.

5. Fixation determination

For this experiment a speed threshold of 12.5° per second was used for determining the beginning and end of saccades. Unfortunately, this did not allow for the detection of extremely short fixations. These occurred typically when the initial saccade was directed towards the distractor. Due to the important influence of these saccades on the response variables, an additional fixation criteria, direction change, was introduced. Eye movement vectors were determined by looking at consecutive gaze locations and computing the vector from one location to the next. Whenever the angle between two consecutive eye movement vectors during a saccade was larger than 60° it was defined as the end of a saccade and beginning of a fixation. If the following gaze recordings dropped below the speed thresholds, they were included into the fixation; otherwise the fixation ended. Visual inspection of scene overlays showed that this method effectively separated saccades from fixations and managed to capture very short fixations that were apparent through a sharp direction change only.

6. Response Variables

In order to assess the contributions and interactions of the top-down, bottom-up, and semantic factors on attention allocation and eye movements, six response variables were analyzed.

The first two variables, namely the number of fixations until target fixation and the time until target fixation, are closely related. Both are indicators of the search performance but also on the capture of overt attention. Longer times or higher numbers show reduced

performance and thus attentional capture by the distractor. The time until target fixation was measured from scene onset until the first fixation lands on the target area. The target area is a rectangle around the target extending 2° or 96 pixels out to either side from the minimum and maximum target-coordinate values in both the x and y direction. The number of fixations until target fixation is counted starting with the first fixation leaving a circle with a radius of 2° around the screen center up until and including the first fixation on the target area. That means, if the first saccade lands on the target area, the number of fixations until target fixation is one.

The reaction time was measured from scene onset until the participants pressed the spacebar to indicate a successful search. Since participants were instructed to first fixate the target and then press the spacebar, this time can not be compared to other search experiments where the reaction time is usually the only response variable and does not require a concurrent fixation. The incentive for this instruction was to discourage participants from making guesses. The measured reaction time is still different from the time until target fixation since participants frequently pressed the spacebar during the saccade onto the target. Therefore, the time at which participants pressed the spacebar is still a valid measure of reaction time.

The next response variable, initial saccade latency, is the time which expires from scene onset until the end of the last fixation within the 2° circle around the screen center. It is a measure of the time spent for covert orienting before a participant performs the first saccade.

The length of the first on-target saccade measures the perceptual span. It indicates how far a target can be from a fixation location and still recognized and saccaded to.

Lastly, the initial saccade direction tells whether the target, distractor or hiding location captured the overtly deployed attention.

C. RESULTS AND DISCUSSION

In this section the results from the five experimental conditions for all response variables will be presented and discussed.

1. Target Only Trials

For the target only trials, the target was placed at eccentricity levels 0, 2, and 3. The saliency levels had been 0, 1, and 3. Eccentricity level 1 and saliency level 2 were not used in this condition. All response variables, except for initial saccade direction, were submitted to an analysis of variance (ANOVA) that examined the influences of factor levels on the response.

a. Number of Fixations Until the First Target Fixation

The average number of fixations until the first target fixation was 1.19 and standard deviation was 0.53, with the majority of fixation numbers being 1 (98 out of 114). The ANOVA did not show any influence of factor levels (p=0.2603). Apparently participants had generally been able to spot the target without having to overtly scan the scene.

b. Time Until the First Target Fixation

The average time that elapsed before the target has been fixated was 893 ms with a standard deviation of 271 ms. Again, the ANOVA did not reveal a difference based on the factor levels (p=0.4042).

c. Initial Saccade Latency

The average latency of the initial saccade was 638 ms with a standard deviation of 279 ms. There was no factor effect on this response variable (p=0.8346).

d. Length of the First On-target Saccade

The length of the first on-target saccade was 492 pixels on average. The standard deviation was 282 pixels. This amounts to 10.25° of visual angle and 5.88°of visual angle respectively. There was a main effect of target eccentricity ($p<0.0001$) with all levels being different from each other as shown by a comparison of each pair using a t-test. This is due to the fact that participants fixated the target after the first saccade. Therefore, the length of the first on-target saccade correlated with the target eccentricity. This is confirmed by looking at the mean length of the first on-target saccades for each of the eccentricity levels. At a target eccentricity of 5° of visual angle the mean first on-target saccade length was 4.58° of visual angle (standard deviation: 3.42°). For eccentricities of 13° and 17° of visual angle, the mean and standard deviation were 12.19°, 2.92° and 14.94°, 5.38° of visual angle respectively. Target saliency did not show an effect on the length of the first on-target saccade ($p=0.7154$).

e. Reaction Time

The response time was 759 ms with a standard deviation of 267 ms. No factor influences were found for this response variable ($p=0.4811$). However, there was a notable difference between the time until first target fixation and response time. This means that participants reported the target found before they fixated on the target, indicating that in the absence of any distracting element, participants identified the target without having to foveate it.

f. Initial Saccade Direction

The first saccade was directed towards the target in almost all of the trials. In only 1 out of 114 trials was the first saccade not directed towards the target. In other words 99.1% of all first saccades were directed towards the target.

g. Discussion

In the absence of a distractor, humans are apparently able to find and fixate the target independent of saliency and eccentricity for not only eccentricity values up to 20° of visual angle, but also for very high as well as very low saliency values. It is important to note that the target only trials were shown together with the target and distractor trials in random sequence. In addition to that, the participants were told that there could also be no targets in the scene. That means that the participants could not rely to fixate into the direction which seemed to contain an object. For all trials, the target had to be identified before saccaded to. The fact that the response time was smaller on average than the elapsed time before the first on-target saccade also shows, that the target had been identified before it had been fixated. This was even the case for the saliency level 0 of the target.

2. Target and Distractor Trials

For the target and distractor trials three individual trials were manually removed from the analysis. The number of fixations until first target fixation of these trials was 7. This was the highest number for this condition with the second highest number being 6. A visual inspection of eye track overlays on the scenes revealed, that the eye tracking was intermittent during the trials. This resulted in spurious gaze tracks with seemingly large direction changes. These direction changes led to erroneous fixation registrations and the trials needed to be discarded. In addition to that, for one trial, visual inspection revealed the miss of a fixation after a saccade going into the direction of the distractor. The fix was apparent due to a sharp direction change but was not detected by the fixation determination algorithm. The trial was encoded manually and added to the trials being analyzed.

In this condition, the eccentricity and saliency levels of the target had been 0, 1, and 3 for both factors. The eccentricity of the distractor was set to 0, 1, and 3, and the saliency could assume levels 0, 2, and 3.

a. Number of Fixations Until the First Target Fixation

With a mean of 1.52 fixations until the first target fixation (σ = 0.77) and 59.1% of fixations landing on the target after the first saccade, there was a strong decrease as compared to the target only trial. A two-sample t-test based on condition showed that this difference was statistically significant (p=0.002). The eccentricity of the target showed a positive main effect on the number of fixations (p=0.0099), that is, the higher the eccentricity the higher the number of fixations. In addition to that, there was a main effect of target saliency (p=0.0219) with a larger saliency resulting in lower numbers of fixation until the first target fixation. Contrary to this effect, saliency level 3 showed a slightly higher mean number of fixations until the first target fixation as compared to level 1, namely 1.44 fixations as opposed to 1.37 fixations (see Figure 8). Comparison of each pair using Student's t showed that this difference was not significant. However, it inidcated, that once the saliency reached a certain level, there was no further benefit with respect to the number of fixations needed to spot the target. Still, the results pertaining to target eccentricity and saliency matched the hypothesis that targets closer to the initial fixation location and targets with higher saliency can be spotted and fixated faster. The effects of distractor eccentricity (p=0.8436) and distractor saliency (p=0.3785) were not significant. However, plots of the data showed an interesting trend for distractor saliency (Figure 8). The average number of fixations before target fixation remained similar for saliency levels 0 and 2 with 1.58 and 1.59 fixations respectively. For saliency level 3, the average number of fixations dropped to 1.40.

b. Time Until the First Target Fixation

The time until the first target fixation showed a similar pattern as the number of fixations until the first target fixation but the differences to the target only trials were less pronounced. The mean for this measure was 966 ms with a standard deviation of 311 ms. There is a main effect of target eccentricity (p=0.0015) and target saliency (p=0.0065) on the time until the first target fixation. Higher eccentricity increased the time and higher

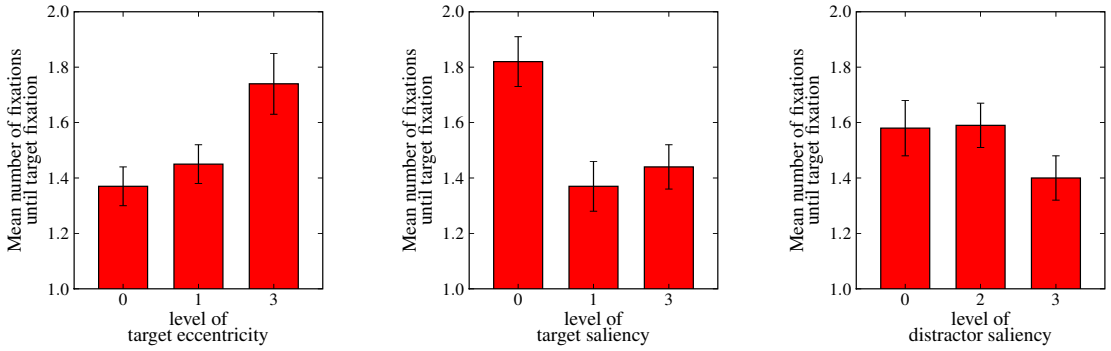

Figure 8: Mean number of fixations until target fixation. From left to right, factor level for target eccentricity, target saliency and distractor saliency.

saliency decreased it (Figure 9). A comparison of each pair of target saliency levels using Student's t-test showed that there was no difference between level 1 and 3, but between all other saliency levels. Similarly, a comparison of each pair of target eccentricity levels using a Student's t-test did not reveal a difference between eccentricity levels 0 and 1 but between all other levels. There were no reliable effects of distractor saliency (p=0.5693) or eccentricity (p=0.9544), nor were there any interactions of factors.

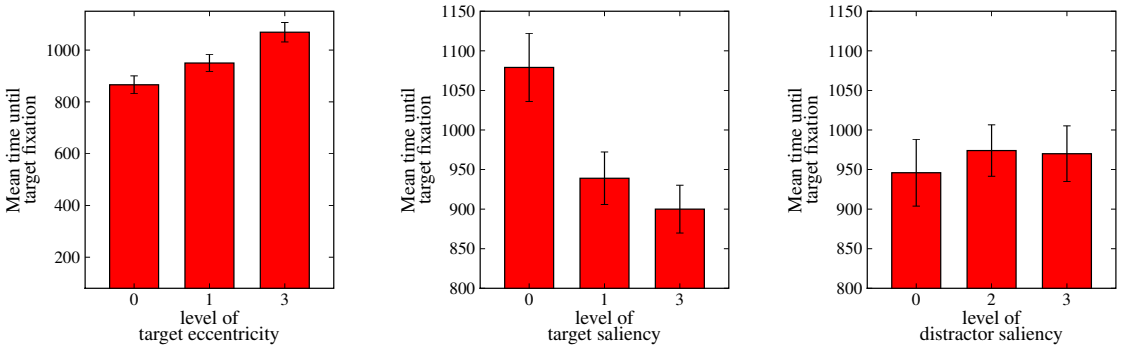

Figure 9: Mean time until target fixation. From left to right, factor level for target eccentricity, target saliency and distractor saliency.

c. Initial Saccade Latency

The initial saccade latency for the target and distractor trials had a mean of 660 ms and a standard deviation of 274 ms. The ANOVA did not show a difference based of the factor levels (p=0.8096).

d. Length of the First On-target Saccade

With a mean of 537 pixels and a standard deviation of 325 pixels the length of the first on-target saccade was only slightly higher as in the target only trials, but there was a large difference in the factor influences. Target eccentricity ($p<0.0001$), target saliency ($p<0.0001$), and distractor saliency ($p<0.0001$) had a main effect. Eccentricity of the distractor ($p=0.7911$) did not show a main effect and there were no interactions. Higher eccentricities of the target increased the size of the initial on-target saccade, which mirrors the observations of the target only trials. Saliency of the target decreased the response with increasing factor levels (Figure 10). An interesting effect could be observed for the saliency of the distractor. The effect of distractor saliency was highest for a saliency level of 2 (mean = 593 pixels). The saccade length was smaller for the other two saliency levels, 0 and 3 with means of 510 pixels and 476 pixels, respectively (Figure 10). A comparison for each pair using a student t-test showed that there was only a significant difference between level 2 and 3 ($p=0.0188$) but not between levels 0 and 2 ($p=0.1357$) or levels 0 and 3 ($p=0.5700$). That means the saccade length was the shortest for the highest level of distractor saliency.

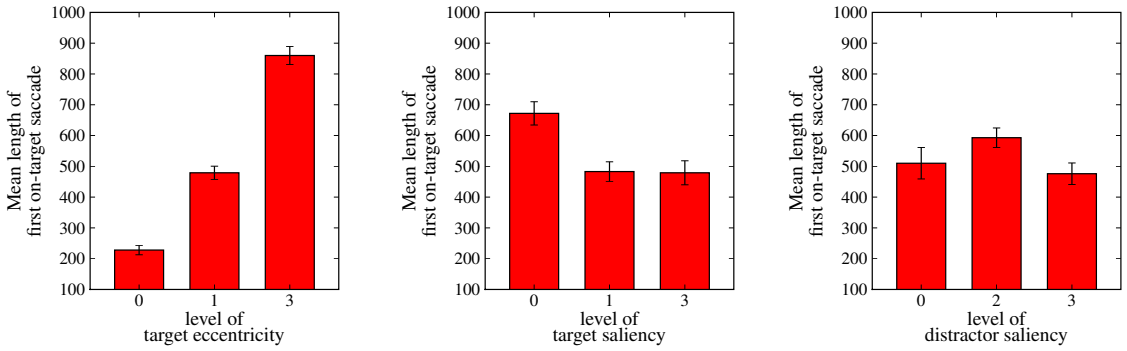

Figure 10: Mean length of first on-target saccade. From left to right, target eccentricity, target saliency and distractor saliency.

e. Reaction Time

The reaction time was less than the time until target fixation. The mean reaction time was 857 ms with a standard deviation of 323 ms. Similar to the time until target detection, the target eccentricity ($p=0.0050$) and the target saliency ($p=0.0016$)

showed a main effect. The reaction time increased with increasing target eccentricity and it decreased with increasing target saliency. However, a Student t-test comparing each pair of target saliency levels showed no difference between level 1 and 3. This means that there was no improve in reaction time when the target saliency was increased from level 1 to level 3. There were no statistically significant effects of distractor saliency (p=0.6536) or eccentricity (p=0.8567).

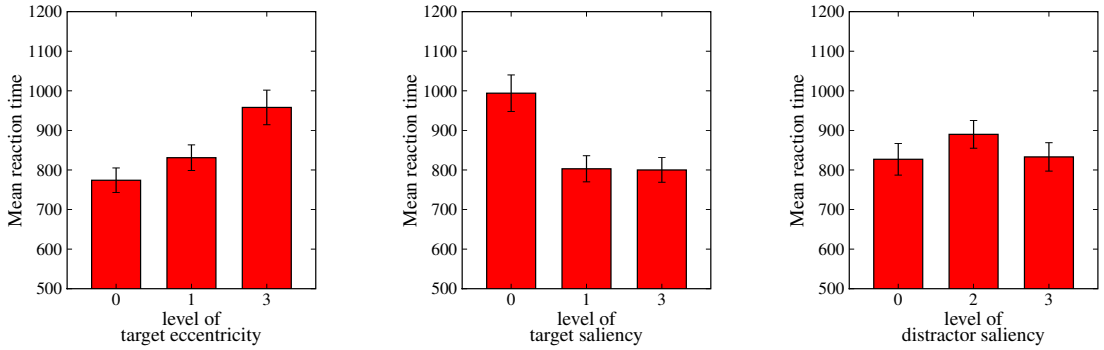

Figure 11: Mean reaction time. From left to right, target eccentricity, target saliency and distractor saliency.

f. Initial Saccade Direction

In 157 of 225 trials the initial saccade was directed towards the target and in 68 trials towards the distractor. This means that in 69.8% of the trials the initial saccade went into the direction of the target. This is considerably less than in the target only trials and a χ^2-test showed statistical significance between the two conditions (p<0.0001). Still, the majority of the fixations went towards the target. That means that even with a distractor the likelihood of a first target fixation is well above chance.

g. Comparison with the Target Only Trials

In order to compare the differences of the response variables between the target only trials and the target-distractor trials, paired t-tests were performed for the number of fixations until first target fixation (p=0.0020), time until first target fixation (p=0.0187), initial saccade latency (p=0.45), length of first on-target saccade (p=0.0149), and response

time (p=0.0288). For the same purpose a χ^2-test were done for the response variable initial saccade direction (p<0.0001). Only trials with target eccentricity 0 and 3 were included in the comparison. This left out all trials with target eccentricity level 2 for the target only trials and target eccentricity level 1 for the target and distractor trials. This was necessary for making a fair comparison of the two conditions. All of the responses were significantly different across the two conditions except for initial saccade latency.

h. Discussion

The results for the target and distractor trials clearly show that the introduction of the distractor changes the eye movement behavior. This is not only because the across-condition comparison of the response variables show this difference, but also because there are influences of factor levels on the response variables number of fixations until target fixation and time until target fixation that had not been observed for the target only trials. Also, the length of first on-target saccade does not depend on the target eccentricity alone anymore, but also on the saliencies of the target and distractor. The only response not being affected by the introduction of the distractor is the initial saccade latency. This variable is considered to indicate the amount of covert orienting before conducting the first saccade (Henderson et al., 1999). Apparently, participants do not seem to expect a benefit from covert orienting, but the first saccade is programed independently of distractor presence or absence. This is in accordance with the claim of Findlay and Gilchrist (2001). Their active vision account states that there is no benefit of covert orienting over eye movements in real world scene viewing since performing eye movements is nearly as effortless as covert orienting and brings the advantage of high resolution foveal vision.

The two response variables directly connected to the eye movement behavior, number of fixations until target fixation and time until target fixation, do not depend on the distractor saliency nor on distractor eccentricity. This means, that the presence of the distractor alone changes the eye movements even in low saliency and high eccentricity conditions. However, looking at the influence of distractor saliency on the number of fixations until target fixation an interesting phenomenon can be observed. There is the tendency that

higher distractor saliency reduces the number of fixations until the target is found. Moreover, the maximum number of fixations is caused by distractor saliency level 2. This effect is even more pronounced for the response time and the length of first on-target saccade. Although the effect is statistically significant only for the length of the first on-target saccade it shows that the maximum distracting capability of the distractor occurs at a medium saliency level and not at the maximum saliency level.

This is surprising because based on the theory of bottom-up attention allocation a more salient distractor should draw the eyes with a higher likelihood. For example, the visual attention model of L. Itti et al. (1998), assigns the first locus of attention to the location with the highest saliency. In the case of a top-down guided search task, this location would still get the attention if it exceeds the task-modulated target saliency (Navalpakkam & Itti, 2005). Non-target locations with less saliency will only get assigned the focus of attention if the inhibition of return mechanism inhibits the previously active locations. On the other hand, the results presented here suggest that the medium salient locations draw the eyes away from the target with a higher likelihood. One explanation for this effect is that a distractor with a high saliency might draw the attention of the observer, but no fixation takes place because there is sufficient information content to allow for peripheral processing. In the case of lower saliency, there is a higher need to process the distractor location foveally to gain sufficient information for making a target or no-target judgment, and thus the eyes follow the focus of attention already allocated to the distractor.

This is supported by the results of the initial saccade latency. Although the plot of initial saccade latency by distractor saliency shows the highest value for saliency level 3, there is no statistically significant influence of distractor saliency on initial saccade latency. This means that the covert attention allocation of saliency levels 2 and 3 does not differ, but in one case the eyes follow the locus of attention and in the other case they do not. On the other hand, a difference in initial saccade latency between the target only and the target and distractor condition could not be observed. Therefore, one has to conclude

that there is no covert attentional capture except for the one preceding the overt attention allocation to the distractor as reported by Hoffman and Subramaniam (1995).

Based on the rather minute influence of distractor factors it is clear that the governing influences on the eye movements are the appearance and location of the target. Targets closer to the initial fixation and targets with higher saliency show a significantly reduced number of fixations and time until target fixation. The mean number of fixations until target fixation grows more than linearly with eccentricity. Still, even at the highest eccentricity level of this experiment can the target be fixated after the first saccade. This indicates that the task influence can override stimulus-driven attentional capture, but a distractor can as well override the task demands. The strength of task influence is reduced with higher eccentricity.

The time until target fixation shows a similar effect, but in this case the effect of eccentricity shows a linear relationship. Since the two metrics are closely related this raises the question as to why eccentricity has a linear effect in one case and a more than linear effect in the other case. Fixations can land either on the background or on the distractor. Especially the fixations that land on the background are extremely short and add very little time as compared to the the total time, whereas the influence of one additional fixation has a large contribution to the overall mean. Some fixations are so short that they could not be detected by a speed threshold, indicating that they are shorter than a single frame. These very short fixations could be frequently observed for initial saccades going towards the distractor. In these cases it seems that the participants had already noticed that the eyes were going in the wrong direction during the saccade and thus the next saccade programming was already taking place before the fixation.

Looking at the influence of saliency on the number of fixations until target fixation and on the time until target fixation it is obvious that saliency does not show a linear relationship with the mentioned metrics. Raising the target saliency from level 1 to 3 does not decrease the number of fixations or the time until target fixation as much as the change from saliency level 0 to 1. In other words, once the saliency exceeds a certain threshold,

a further increase does not speed up the search process. In the presented experiment the threshold turned out to be somewhere between saliency level 0 and 1. Paired t-tests for the time and number of fixations until target fixation showed a statistical significant difference between level 0 and 1 but not between 1 and 3.

Despite the apparent influence of the distractor, there is still a very high number of trials in which targets were fixated after the first saccade even for targets with high eccentricity and low saliency. This means that a distractor can always draw the attention but it does not have to. Also, it is very interesting that the distractor appearance is statistically significant for the length of the first on-target saccade only. Apparently, the bare presence of the distractor is sufficient to distract from the task. A similar effect is reported by Born and Kerzel (2008). They showed that the initial saccade latency changed with the introduction of a distractor. This is at odds with the observations of this study, which did not show a change in initial saccade latency between the target only condition and the target and distractor condition. However, there is a large difference between this experiment and the one of Born and Kerzel. Their targets and distractors were upright Gábor patches and there was not a specific search task. Possibly, the task dependence of this experiment overrode the effects on the initial saccade latency that had been reported by Born and Kerzel. They already suspected that a task influence could potentially change the effects they had observed.

In contrast to the findings presented here, Theeuwes (2004) argued that task demands can not override attentional capture. The experimental paradigm of Theeuwes as well as the reported results pertain to covert attention allocation, which we did not examine directly. However, the response variable initial saccade latency is a measure of the time spent for covert attention allocation before the first saccade. Although the initial saccade latency was higher in the distractor condition as compared to the target only condition, the difference was not statistically significant. This means that there is no evidence for attentional capture overriding task demands, which is in accordance with the findings of Bacon and Egeth (1994) and Einhäuser, Rutishauser, and Koch (2008) who both showed

that task demands can take precedence over stimulus-driven attentional capture. In addition to that an influence of the target and distractor saliency on the initial saccade latency could not be observed. According to Theeuwes (2004), Bacon and Egeth (1994) failed to observe attentional capture due to a lack of target saliency. In the experiment presented here, this was not the case, but still it was not possible to observe covert attentional capture before the first saccade either.

What could be seen was overt attentional capture that could be overridden by the task demands, depending on target eccentricity and saliency. The distractor can, but does not have to, capture the attention. This occurs mostly if the target has a low contrast or is placed at a high eccentricity. This means that with better target visibility, the distractor effects get smaller. There are several reasons for these contrasting results. Most likely, the cause of this difference lies in the different quality of the target and distractor. Theeuwes (2004) used geometric shapes whereas the experiment presented here used actual real world objects. One of these, the human figure,is a Gestalt, which plays a very special role for humans in general. Detecting and interacting with other humans are acts people are engaged in almost every minute in their lives. Therefore, it seems to be reasonable to assume that the human Gestalt features have a stronger effect than basic geometric shapes. It is hardly conceivable to see a human figure as a shape singleton, which was the case in the experiment of Theeuwes (2004). This shows, that one has to be careful extrapolating results of visual search experiments with artificial stimuli to search in realistic scenes.

In summary, it can be seen that a clearly visible human target within a certain range can be spotted fast, and it is hard for a visually salient distractor to reverse the top-down effects. However, if the information content for top-down processing lessens, there is a larger likelihood that the distractor captures attention. Thus, for a scenario that would not provide any semantically relevant information, an eye movement model should first determine whether or with which probability the target could be found based on its eccentricity and saliency. And only if the target can not be found should the eye movement

be drawn to the salient location that attracts the gaze most, which is not necessarily the location with the maximum saliency.

Determination of the saliency should also consider top-down guidance or top-down modulation as does, for example, the Guided Search Model (Wolfe, 1994) . Also, Pomplun (2006) has recently shown the influence of top-down modulation on search in naturalistic scenes. However, neither guided search nor the findings of Pomplun can explain how search can be guided to a target for which the exact appearance is not known beforehand. This means that the top-down guidance needs to rely on more general features of the search target, which in this study is a human figure. In addition to that, it is assumed, that semantically relevant information is guiding the eyes to promising locations. Therefore we designed the second part of our experiment looking at the influence of a location which would serve as a semantically guiding cue.

3. Target and Hiding Location Trials

In order to examine the influence of a semantically informative object on a visual search task, the experiment comprises a search condition in which stimuli containing a human target and a hiding location, but no visually salient distractors, are present. The hiding location, namely a doorway in the background wall, serves as semantically relevant location. This is a location associated with human target presence. In some sense, the doorway can be considered a distractor since it does not help the search in this experimental condition, but will be misleading if it is perceived as semantically guiding. However, it is distinct from a purely visual distractor, which can not provide guidance for a successful search in any case. The expectation is to see different eye movement behavior in this condition due to the semantic salience of the hiding location.

From the captured data a total of 4 trials were excluded manually. Due to eye-tracking error, three of those were showing eye tracks not consistent with the eye tracks observed for the other trials. The eye tracks of the excluded trials were located considerably above or below the horizontal line along which the search entities were placed, but with sporadic fixation landing close to the target area. This resulted in very long search times

and fixation numbers until target fixation, which are not representative of the overall data. Another trial was excluded because the reaction time amounted to 8008 ms. All trials had been restricted to a search time of 5 sec. This restriction was violated and hence the trial was discarded.

The factor levels in this condition were 0 and 1 for the target saliency, 2 and 3 for the target eccentricity and 0 and 3 for the hiding location eccentricity. No attempt was made to vary the visual saliency of the hiding location. The level was fixed to a medium saliency such that it would be easy to recognize, but not draw the attention due to high visual conspicuity.

a. Number of Fixations Until the First Target Fixation

The mean number of fixations until target fixation was 2.02 with a standard deviation of 1.58. The target was fixated after the first saccade in 49 out of 95 trials (51.6%). The number of fixations until target fixation showed a main effect of target saliency ($p<0.0001$) with higher saliency resulting in fewer fixations. The other two factors did not show a reliable effect. Although not statistically significant, it is interesting to note that the number of fixations increased with higher hiding location eccentricity. This indicates, that the target detection performance is reduced with higher eccentricity of the semantically relevant location.

b. Time Until Target Fixation

The time until target fixation had a mean of 1118 ms and a standard deviation of 429 ms. Again, the target saliency was the only factor that showed an effect ($p<0.001$) with higher saliency, which resulted in less time until target fixation. Higher eccentricity of the hiding location, on the other hand, results in an increase of the time until target fixation, but this effect is not statistically significant. However, it strengthens the identical observation of this effect for the number of fixations until target fixation.

c. Initial Saccade Latency

With a mean of 605 ms and standard deviation of 286 ms, the initial saccade latency was in the same range as in the previous two conditions. Similarly to the previous two conditions, the initial saccade latency did not show any effect of the factors.

d. Length of the First On-target Saccade

The length of the first on-target saccade had a mean of 743 ms with a standard deviation of 325 ms. In contrast to the previous two conditions, and rather surprisingly, the length of the first-on target saccade did not show an effect of target eccentricity ($p=0.1700$). In fact, none of the factors showed an effect.

e. Initial Saccade Direction

Fifty five initial saccades were directed towards the target (57.9%) and 40 towards the hiding location (42.1%). A χ^2-test showed a significant influence of target saliency ($p<0.0001$), but no effect of any other factor.

f. Reaction Time

The mean reaction time was 1036 ms with a standard deviation of 619 ms. The response was dependent on target saliency ($p<0.0001$), but neither on target nor on hiding location eccentricity. The reaction time was indirectly proportional to target saliency.

g. Comparison with the Target Only and the Target and Distractor Trials

In order to compare the target and hiding location responses with the response of the previous two conditions, t-tests were performed. From the two conditions that were compared, only trials with matching factor levels were included. That means that all trials containing a factor level not being covered in the comparison condition were excluded from the comparison. The eccentricity levels in the target only condition, for example, were 0, 2, and 3, whereas in the target and hiding location condition they were 2 and 3 only. In this case, all trials with a target eccentricity level of 0 were not included in

the comparison. With respect to the hiding location saliency, as compared to the distractor saliency, all trials with distractor saliency levels of 2 and 3 were included.

Comparison of the target and distractor condition with the target and hiding location condition, using a χ^2-test for initial saccade direction and t-tests for all other response variables, did not show any differences in the responses across condition.

h. Discussion

The direct comparison of the target and distractor condition and the target and hiding location condition did not show any differences with respect to the response variables. Superficially, the overall means in the two conditions seem to strongly differ, but this is due to a different set of factor levels being used in the two conditions. The target saliency in particular was varied across level 0 and level 1 in the target and hiding location condition, but in the previous condition, level 3 was also included. With a target saliency level of 3, the search performance was extraordinarily good, and thus the overall means of the responses were better. In order to not confound the comparison results, the comparison of the two conditions had to include only trials with common factor levels. This comparison did not show statistical significant differences for any of the employed response variables. A similar comparison with the target only conditions showed that all responses of the target and hiding location condition were significantly worse. This is the expected result since the hiding location did not help find the target, but was actually misleading due to its locus in the opposite hemifield of the target.

Based on these two results it seems that the search performance of the target and distractor condition and the target and hiding location condition was the same. This could be interpreted in two different ways. First, it could be that the doorway just served as a visually salient distractor and was not perceived as a location with semantic informativeness. The second possible explanation would be that despite being perceived as a semantically relevant location, the hiding location had the same effect as the distractor. This would mean, that a visually salient distractor and a semantically salient distractor would elicit the same eye movement behavior.

However, both of these hypotheses are in contrast to the factor influences exhibited in the target and hiding location condition on the one hand, and the target and distractor condition on the other hand. Looking at the factor influences of the two conditions, it is most notable that the eccentricity of the target did not influence any response variable in the target and hiding location condition, whereas it influenced all of the responses except for initial saccade latency in the target and distractor condition.

Before discussing this phenomenon, a look at the effect of target saliency is in order since target saliency was the most influential factor on the response variables in the target and hiding location condition. The number of fixations until first target fixation and the time until target fixation decreased as target saliency increased. Although this effect could be observed in the target and distractor condition as well, it is very interesting, that the decrease of the responses with increasing target saliency was much higher in the target and hiding location condition than in the target and distractor condition. The average time until target fixation in the target and distractor condition was 939 ms for saliency level 1 and 1079 ms for saliency level 0 which amounts to a difference of 140 ms. In the target and hiding location condition, this difference was 361 ms. For saliency level 1 the time until target fixation was 940 ms and for saliency level 0 it was 1301 ms. This phenomenon was even more pronounced for the number of fixations until target fixation. In the target and distractor condition the mean number of fixations until the first target fixation was 1.37 for target saliency level 1 and 1.81 for target saliency level 0, a difference of 0.44 fixations. The difference in the target and hiding location condition was 1.60 fixations with means of 2.83 and 1.23 fixations for target saliency levels 1 and 0 respectively (see Table 1).

This means that in the case of the higher target saliency, the effect of the distractor on capturing the eyes was at least the same, and possibly a little higher than the effect of the hiding location. The distractor was distracting more than the hiding location in this case. This is in contrast to the low target-saliency case. There, the hiding-location captured the eyes more strongly than the distractor, which means that the distractor distracted less than the hiding location. Analogous effects could be observed for the response

	target and distractor		target and hiding location	
	saliency target		saliency target	
response variable	level 0	level 1	level 0	level 1
average time until first target fixation	1079 ms	939 ms	1301 ms	940 ms
average number of fixations until first target fixation	1.81	1.37	2.83	1.23
proportion of initial saccades directed towards the target	0.35	0.86	0.26	0.90
average reaction time	994 ms	803 ms	1325 ms	771 ms

Table 1: Comparison of the responses of the target and distractor condition with the responses of the target and hiding location condition by target saliency level.

variables reaction time and initial saccade direction (see Table 1). This shows very clearly that the influence of the distractor and hiding location on eye movements is different in nature.

The explanation for this phenomenon is the semantic informativeness of the hiding location. Apparently, when the target was hard to spot, the eyes were drawn to the hiding location since the target presence was associated with the doorway. Subsequently, that location was inspected for target presence which took time, and additional fixations were produced during this process. Figure 12 shows typical examples of fixations on a hiding location and on a distractor respectively.

This account is further supported by another effect. Although statistically not significant, the eccentricity of the hiding location reduced search performance as shown by the time until first target fixation, number of fixations until first target fixation, and reaction time. All three responses increased as hiding location eccentricity increased. This is in contrast to the effect of the distractor eccentricity in the previous condition. There, the distracting effect was reduced as distractor eccentricity increased (see Figure 13).

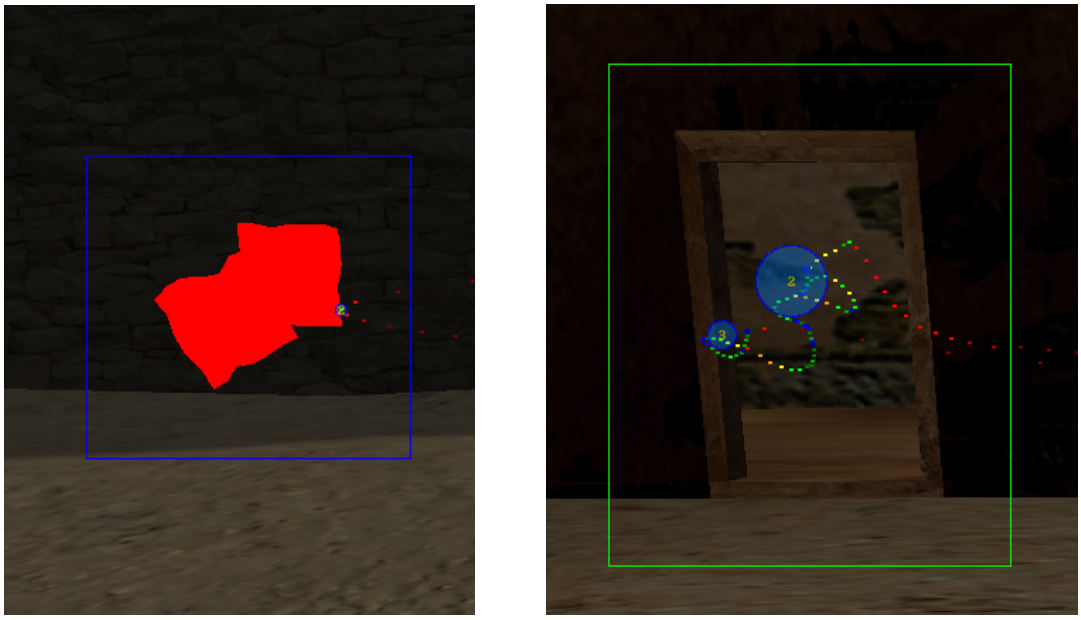

Figure 12: Typical fixations on the distractor and hiding location. Blue circles represent fixations and circle sizes indicate fixation duration. Please note that the distractor was not displayed in this red color during the experiment.

Still, there remains the question why the target eccentricity did not have an effect on any of the response variables anymore, even though there was a reliable, statistically significant, and at least linear influence of target eccentricity in the target and distractor condition. The target eccentricity apparently did not have an influence on whether the target or the hiding location was picked up by the eyes, but it had an influence on whether the distractor or the target was picked up in the previous condition. This means that search performance depended on the target eccentricity in the presence of a visually salient distractor due to its influence on the reflexive control of eye movements. The fact that search performance did not depend on target detection performance in the presence of the hiding location is further evidence for the hiding location capturing the eyes not due to being visually salient but due to being semantically salient with respect to the search task.

Overall, it can be seen that the search performance in the target and hiding location condition was slightly reduced as compared to the target and distractor condition. In addition to that, and more importantly, it was evident that the hiding location had a

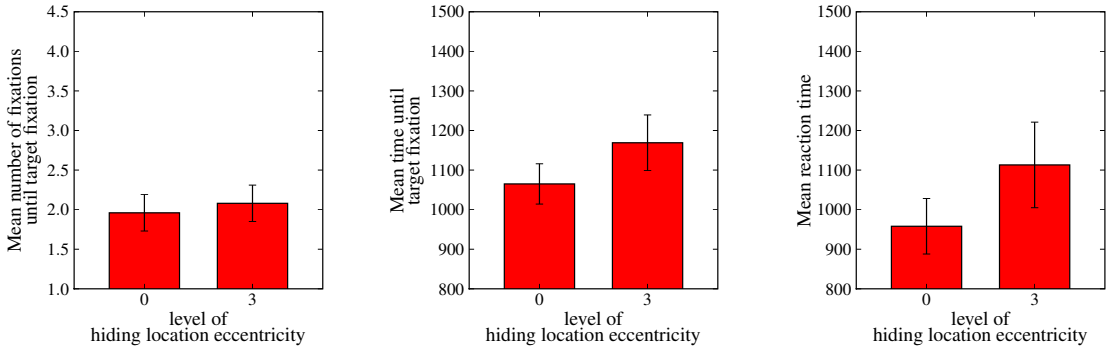

Figure 13: Effect of hiding location eccentricity on the average number of fixations until target fixation, average time until target fixation, and average reaction time.

different effect on the response variables than the distractor. Therefore, it is concluded that this was the result of a task-dependent, non-reflexive guidance of eye movements due to the presence of a semantically informative location.

4. Target and Hiding Location/Distractor Trials

Further assessing the influences of distractors and semantically relevant locations on the eye movements in visual search for a human figure, another condition was examined. In this condition, stimuli containing the target in one hemifield and a visually salient distractor as well as a hiding location in the opposite hemifield were shown to participants. The same visual distractor and the same doorway as in the previous conditions were used. The factor levels were set to level 0 and level 1 for the target saliency and to level 2 and level 3 for the target eccentricity. Distractor saliency was varied between levels 1 and 3 and distractor eccentricity between levels 0 and 2. The hiding location eccentricity could assume level 0 or level 2.

A total of 9 trials had to be manually excluded from analysis. Visual inspection revealed that eye-tracking was unstable during these trials, which resulted in extraneous fixations. In addition to that, the results of 2 trials were manually adjusted. In these trials, a fixation that was apparent through a strong change in direction, was not detected by the fixation detection mechanism. The number of fixations was increased by one and the initial saccade direction needed to be adjusted.

a. Number of Fixations Until Target Fixation

The number of fixations until target fixation had a mean of 2.35 fixations with a standard deviation of 2.01 fixations. Only in 160 out of 405 trials (39.3%) was the target fixated after the first saccade, and in 135 trials (33.3%) the target was fixated after the second saccade. This was a strong, statistically significant increase (p=0.0454) in the mean number of fixations until target fixation as compared to the target and hiding location condition, indicating an even further decreased target detection performance.

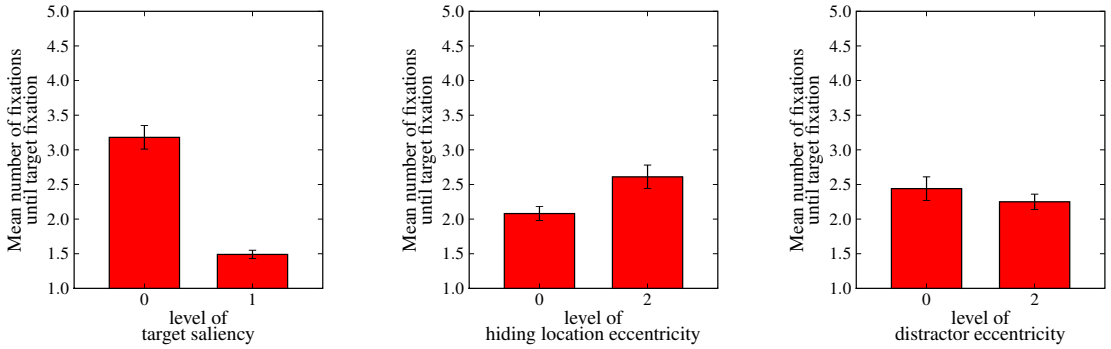

Figure 14: Effects of target saliency, hiding location eccentricity, and distractor eccentricity on the number of fixations until target fixation.

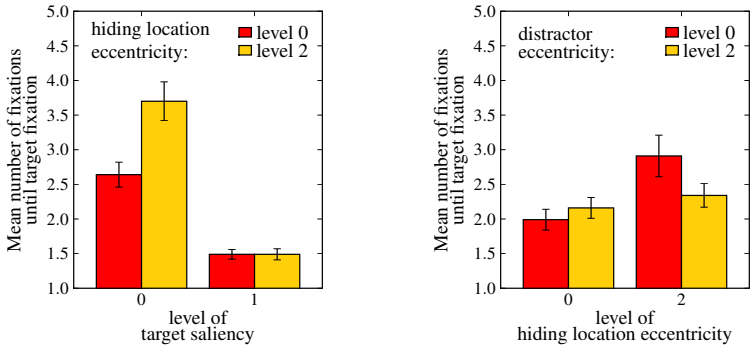

Figure 15: Interaction effects of target saliency with hiding location eccentricity, and hiding location eccentricity with distractor eccentricity on the number of fixations until target fixation.

There was a main effect of target saliency (p=0.0002) and hiding location eccentricity (p=0.0002) on the number of fixations until target fixation. Higher target saliency resulted in a lower number of fixations. This outcome is in accordance with the

observation of the previous conditions, as targets are faster to spot once their saliency or contrast reaches a certain level. Increasing hiding location eccentricity on the other hand increased the number of fixations until target fixation. This interesting effect was already observed in the target and hiding location condition, but there it was not significant. In addition to the main effects, there was also an interaction between target saliency and eccentricity of the hiding location. Inspection of a plot of the number of fixations until target fixation against hiding location eccentricity grouped by target saliency showed that increasing hiding location eccentricity increased the number of fixations only in the case of low target saliency (see Figure 15). There was also an interaction between distractor eccentricity and hiding location eccentricity. The increasing effect of hiding location eccentricity on the number of fixations until target fixation was modulated by the distractor eccentricity. Higher eccentricity of the distractor reduced the effect of hiding location eccentricity (see Figure 15).

b. Time Until Target Fixation

The mean time until target fixation in the target and hiding location/distractor condition was 1209 ms, which is the maximum time observed so far. The standard deviation was 491 ms. The time until target detection showed main effects of target saliency ($p=0.0001$) and hiding location eccentricity ($p<0.0001$), as well as an interaction of these two factors ($p=0.0004$) as can be seen in Figures 16 and 17. As target saliency increased the time until target fixation decreased, and as eccentricity increased the time until target fixation increased. However, the effect of increased time until target detection with increased hiding location eccentricity vanished for target saliency level 1. This means that at target saliency level 1, the target was easy to spot and the hiding location had no importance for the search task. In addition to that, there was an interaction of distractor saliency and hiding location eccentricity ($p=0.0372$). The increase of time until target fixation caused by the increase of hiding location eccentricity was less as distractor saliency decreased.

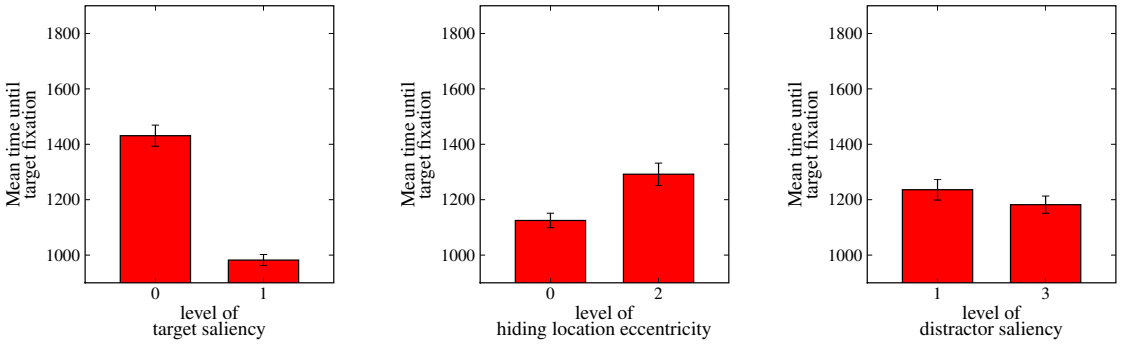

Figure 16: Effects of target saliency, hiding location eccentricity, and distractor saliency on the time until target fixation.

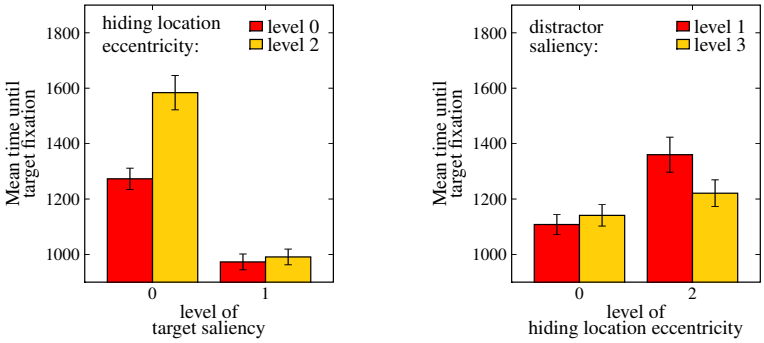

Figure 17: Interaction effects of target saliency with hiding location eccentricity and distractor saliency with hiding location eccentricity on the time until target fixation.

c. Initial Saccade Latency

With a mean of 640 ms and a standard deviation of 290 ms, the initial saccade latency was in the same range as observed in the previous conditions. Similar to the previous conditions, there was no factor effect on the initial saccade latency.

d. Length of First On-target Saccade

The length of the first on-target saccade showed a similar size as in the target and hiding location condition. The mean length was 760 pixels with a standard deviation of 362 pixels. This amounts to 15.8° of visual angle and 7.5° of visual angle respectively. No significant factor effect was observed for the length of the first on-target saccade.

e. Initial Saccade Direction

The initial saccade was directed towards the target hemifield in 45.7% of the trials. This was much less than in all of the previous conditions, indicating a larger distracting effect of the distractor and hiding location together than on their own. A χ^2-test of initial saccade direction revealed a main effect of target saliency ($p<0.0001$) and a main effect of hiding location eccentricity ($p=0.0319$). Increasing target saliency increased the number of initial saccades directed towards the target. Conversely, the number of initial saccades towards the target decreased with increasing hiding location eccentricity (Figure 18).

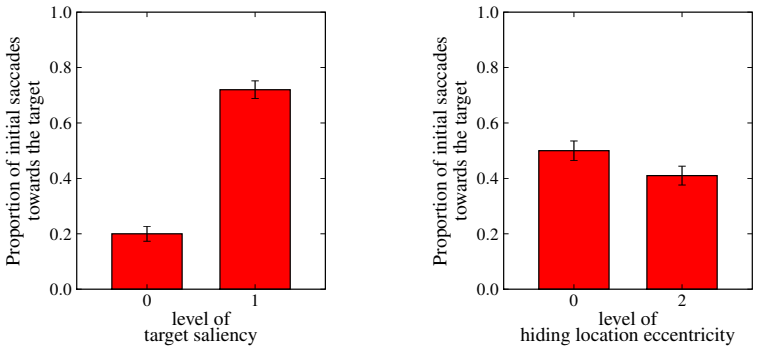

Figure 18: Effects of target saliency and hiding location eccentricity on initial saccade direction. The graphs show the ratio of initial saccade being directed towards the target.

f. Reaction Time

Similar to all the other responses, the reaction time was also further increased. The mean was 1085 ms and the standard deviation 555 ms. Similar to the time until first target fixation, the reaction time showed main effects of target saliency ($p=0.002$) and hiding location eccentricity ($p<0.0001$), as well as an interaction of these two factors ($p<0.0001$). Again, increasing target saliency decreased the reaction time and increasing hiding location eccentricity increased reaction time. This effect was almost non-existent for target saliency level 1 (Figure 19).

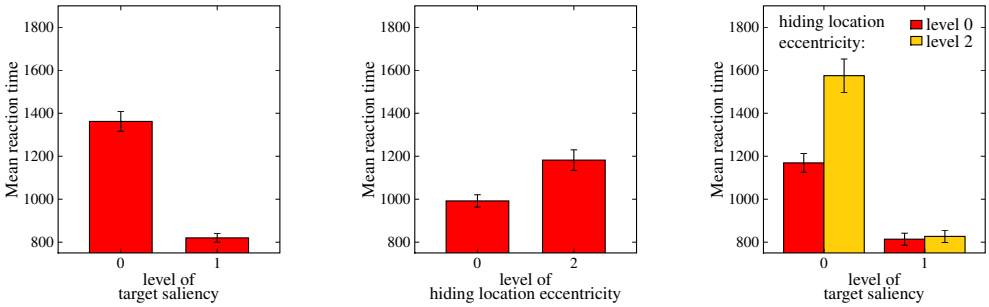

Figure 19: Main effects of target saliency, hiding location eccentricity and interaction effect of target saliency and hiding location eccentricity on reaction time.

g. Comparison with the Target and Distractor Condition

A comparison of the target and distractor condition with the target and hiding location/distractor condition using t-tests for the number of fixations until target fixation ($p<0.0001$), time until target fixation ($p<0.0001$), initial saccade latency ($p=0.2080$), length of first on target saccade ($p=0.0008$), and reaction time ($p<0.0001$), as well as a χ^2-test for initial saccade direction ($p=0.0009$) showed statistically significant differences for all response variables except for initial saccade latency.

h. Comparison with the Target and Hiding Location Condition

In order to compare the responses of the target and hiding location/distractor condition with the target and hiding location condition, t-tests were performed for the number of fixations until target fixation ($p=0.0454$), time until target fixation ($p=0.0365$), initial saccade latency ($p=0.1412$), length of first on target saccade ($p=0.3284$), and reaction time ($p=0.2427$). A χ^2-test was performed for initial saccade direction ($p=0.0318$). The number of fixations until target fixation and the time until target fixation were significantly larger for the target and hiding location/distractor condition, and the number of initial saccades towards the distractor was significantly lower.

i. Discussion

The results of the target and hiding location/distractor trials emphasize the observations of the target and hiding location trials with respect to the roles of the target and

distractor. In addition, there were also interesting interactions between the hiding location and the distractor.

The effects observed for the distractor and the hiding location are very similar to the effects observed in the target and distractor condition and in the target and hiding location condition, respectively. Increasing distractor saliency as well as increasing distractor eccentricity resulted in less fixations and less time until the target was fixated. These response variables, on the other hand, increased with increasing hiding location eccentricity. However, none of the distractor factors had a statistically significant main influence on either response variable. Their significant effects were through interactions with the hiding location eccentricity. The time until target fixation as well as the number of fixations until target fixation increased with increasing hiding location eccentricity. These increases differed depending on the distractor factors. The increase of the time until target fixation with increasing hiding location eccentricity was much stronger when the distractor saliency was lower (see Figure 17).

Similarly, the increase in the number of fixations until target fixation with increasing hiding location eccentricity was much stronger when the distractor eccentricity was lower. This means that a higher distractor saliency improved the detection performance, as already observed in the target and distractor condition. However, this improvement cannot be observed directly by looking at the effect of distractor saliency. Rather, it becomes apparent through decreasing the effect of hiding location eccentricity. These combined effects of the distractor and hiding location also show that a visually salient distractor and a semantically relevant scene location have different effects on the eye movement behavior during a target search. Whereas the influence of a distractor was reduced with higher eccentricity, the influence of the hiding location strengthened. This is an indication for a different level of processing of the two. The visually salient distractor is capturing reflexively controlled attention, and this effect apparently wears off at higher eccentricities. The hiding location on the other hand did not show this reduction. On the contrary, it showed an increase, and therefore it can be concluded that it does not capture the attention based

on reflexive control but based on higher level cognitive processes. Furthermore, the interaction effects of target and hiding location clearly show that both entities affect the search behavior, and neither of them can be excluded from an eye movement model.

Since the time until target fixation and the number of fixations until target fixation are closely related, it is rather surprising that one of these response variables, the time until target fixation, was affected by the distractor saliency, but the other response variable, the number of fixations until target fixation was affected by distractor eccentricity. There is no obvious reason for the two distractor factors affecting these two related responses in a different way. The explanation for this may be that the effects of the distractor showed up in slightly different ways because the distractor did not have a very strong influence among the hiding location. Still, distractors have an effect on the eye movements of an observer. This indicates that the search depends on the properties of a distractor that is displayed together with a hiding location, even if the effects of the distractor are less pronounced than the effects of the hiding location.

Another outcome deserving discussion is the fact that the distractor factors did not have main effects but showed up in interactions only. In the target and distractor condition, there were main effects of distractor eccentricity and distractor saliency on the time until target fixation and the number of fixations until target fixation. These effects disappeared in the target and hiding location/distractor condition. This indicates that the influence of the hiding location on the search performance was stronger than the influence of the distractor. This is also supported by the fact that the distractor factors neither played a role for the reaction time nor for the initial saccade direction. The initial saccade direction is the measure that indicates the number of trials in which the initial saccade was directed towards the target, and conversely the number of trials in which the initial saccade was directed away from the target; that is when the distractor and hiding location did capture the eyes. Since this number was not influenced by the distractor factor, it was apparently the hiding location which was governing the capture of the overtly deployed attention with the distractor playing a minor role only. On the other hand, it is clear that the presence

of the distractor decreased search performance since the ratio of initial saccades being directed to the target was less in the target and hiding location/distractor condition than in the target and hiding location condition. This means that the effects of the distractor and hiding location add up in their capability of capturing the eyes.

The experimental design used here is an indirect approach of showing the influence of semantically relevant locations on the search task. It was not examined whether the search performance actually improved when the target was in a semantically relevant location. However, this indirect approach makes the significance of the semantically relevant information visible. It is important to note that participants were not provided with any information telling them that the target should be expected at the hiding location, nor did they receive any training from which they could have learned this association. The association must have been natural to the participants based on either past experience or a general association of doorways and target presence. This means that semantically relevant information can be accessed for any kind of search, given that there exists such semantically relevant information pertaining to the search task.

This result is in contrast to the findings of Kunar, Flusberg, Horowitz, and Wolfe (2007). They claim that there is no improvement on search performance based on repeated presentations of search arrays with the same arrangement as has been shown by Chun and Jiang (1998), but that it is rather an improvement of response selection. For the presented experiment, this possibility can be ruled out since search benefits due to contextual guidance were not measured. For the same reason, it cannot be concluded that the presence of a semantically relevant cue will actually speed up the search process, but what can be concluded is that examining locations associated with target presence is a human search strategy if the target location is not apparent.

It is not surprising that repeated presentations of search arrays do not improve search efficiency. It is hardly conceivable to interpret search array layouts as a meaningful cue. Apparently, there needs to be semantically relevant information content in order

to provide effective contextual cueing. The hiding location in this experiment showed this property and therefore it could serve as a semantically relevant distractor.

Similar to the findings presented in this work, Brockmole et al. (2006) show that search in naturalistic scenes is facilitated through recurring global and local context, with global context being more influential. However, their findings are based on contexts learned for specific scenes only. The results presented here on the other hand, show that eye movement guidance through semantically relevant information is not constrained to specific pre-learned scene arrangements, but rather relies on stored associations that provide contextual cueing. In other words, the guidance of semantically relevant locations is different in nature as compared to the contextual guidance. Contextual guidance seems to apply to learned co-occurrences of objects only. whereas the guidance of semantically relevant scene locations is based on the meaning of these locations for the current task.

Summarizing the results of the target and hiding location/distractor condition, it can be seen that the general search performance was reduced as compared to the previous condition in which either the distractor or the hiding location were present, which means that the distracting effect of the distractor and hiding location added up. The distractor and the hiding location still showed their general effects, but the negative effects of increasing hiding location eccentricity on the search performance were statistically significant, and the distractor effects were significant in the form of interactions only. This clearly shows, that the search for a real target is strongly influenced by semantic information providing cues to locations with likely target presence but also, even if to a lesser degree, by a visually salient distractor presented simultaneously. In addition, the results show, that a semantic distractor has a different influence on the search task than a visually salient distractor.

D. CONCLUSION

Overall, it can be seen that in a visual search task for a human target, observers can find the target very quickly. Observers hardly ever waste even a single fixation if no dis-

tracting items are present, even if the target has a very low contrast to the background. In the presence of a distractor, either visually salient or semantically salient, search performance is reduced. The search performance as well as the eye movement behavior are influenced by a visually salient distractor as well as by a semantically relevant scene location. The observed influence of the two is not the same; in fact it differs considerably.

The distractor apparently has the capability of drawing the eyes and reducing the search performance, as was shown by the results of the target and distractor condition. On the other hand, it is also clear that the search task can override the overtly attentional capture, that is the capture of the eyes, since a large percentage of initial fixations are directed towards the target, even if the target saliency level is very low. The distractor attracts the gaze less if its eccentricity from the initial fixation location gets larger. Very interestingly, the maximum distracting capability is not tied to the maximum saliency; rather it already has a maximum at an intermediate saliency level with no further increase, even with a possible decrease thereafter. This observation is very important due to the implications it has on the usage of saliency maps for models of attention allocation.

The hiding location draws the eyes of a human observer as well, but as opposed to the distractor, the distracting effect of the hiding location gets larger with increasing eccentricity. This clearly shows that the hiding location attracts the eyes in a different way than the distractor. Most likely this is due to a semantically relevant location being processed differently by the human brain. It is also very apparent that the effect of the hiding location strongly depends on the visibility of the target. If the target is easy to spot, the hiding location plays almost no role. Clearly, the hiding location is processed only if the target is hard to detect. In this case, the eyes are guided to the scene location, which, based on the observers expectations, has the highest probability for finding the target. In the setup of the experiment presented here, this is the hiding location. This means that human observers process the semantic information content, the meaning of scene objects or scene locations, and process this information to help guide the eyes to the target.

In contrast to the results of this work, previous research (Kunar et al., 2007) did not find strong evidence for contextual cueing in visual search. This is apparently due to the fact that they tried to achieve contextual cueing by repeatedly presenting search arrays with the same object arrangements. These repeated arrangements, apparently not allowing the extraction of meaningful context, did not improve the search performance as measured by the reaction time of participants. On the other hand, Brockmole and Henderson (2006) could observe an improvement of search performance for abstract search objects superimposed on naturalistic scenes. However, similar to Kunar et al. (2007) they also had to repeatedly present a given scene layout in order to allow for the participants to learn the target locations in particular scenes. Similarly, Brockmole et al. (2006) found contextual cueing effects for learned scenes with realistic objects in naturalistic scenes.

The results presented in this work significantly extend these findings, showing that humans do not have to learn scene configurations to benefit from contextual cueing. It becomes apparent that meaningful context being associated with the target presence is generally available to humans while performing a realistic search task. The semantics of scene locations and objects are used to inform the search in order to quickly and accurately find the search target. The meaning associated with certain scene locations differs from the contextual cueing observed in previous research, which is based on the co-occurrence of objects.

What are the consequences of the presented results for a model of eye movements in visual search? First of all, there seems to be an order of precedence. If the target is clearly visible, it is almost certain that the target will be fixated directly. If the target is not so clearly visible, locations with semantic relevance for the search task will be inspected with the expectation of finding the target at these locations. If present, a distractor always has a chance to draw the overtly deployed attention. The likelihood for a distractor to draw the eyes is reduced with the distance from the fixation location and with increasing target saliency. This means a model first of all needs to take target presence into account. If the target is not clearly visible, semantically relevant locations are fixation locations to be

generated. Then the distractor draws the eyes with a certain likelihood, being modulated by the distractor distance from the current fixation location and by target saliency.

However, for the modeling approach described in the following chapter, there will be an emphasis on capturing the semantically relevant scene locations from the simulation environment and comparing the predictive capabilities of that information with actual eye-tracking data. This data is collected in an experiment in which participants search for enemy ground soldiers in realistic scenes that depict an urban environment. This experiment, the experimental results, as well as the modeling efforts and the comparison of the model output with the eye tracking data is described in the following chapter.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. PREDICTING EYE FIXATIONS

The previous chapter described a visual search experiment for a human target in simple scenes, being designed to control for salience and eccentricity of the presented objects. The results of the experiment show that top-down factors of semantically relevant scene locations, as well as bottom-up factors, influence eye movements.

This means that an eye movement model needs to include bottom-up information as well as top-down information. The creation of bottom-up salience maps for eye movement allocation was described previously (L. Itti et al., 1998). For this work a re-implementation of this model is used to create bottom-up salience maps. Similarly to these salience maps, relevance maps are created. These maps contain information about locations that are relevant for the search task by highlighting possible hiding locations and regions in which targets blend in well with the background.

The predictive power of these maps is assessed by comparing them with the eye movements of humans looking for human enemy targets in realistic scenes. These eye movements were recorded in an experiment which was part of the experiment described previously.

This chapter will first illustrate the generation of the salience and relevance maps of the scenes, and then the search experiment will be described. Finally, the predictive power of the relevance maps as well as the salience maps of the presented scenes is assessed through comparison of these maps with human fixations on the scenes.

A. COMPUTATION OF SALIENCE AND RELEVANCE MAPS

1. Salience Maps

The computation of the bottom-up maps is, with a few exceptions, performed as described in section II.C.1 on page 33. The differences in the implementation are related to the computation of the intensity maps and the color maps, and also with respect to

the normalization scheme. This section will only describe the differences. A detailed explanation of the model can be found in section II.C.1 on page 33.

The primary intensity map in the Itti model is created by equally weighting the RGB color values of the input image for each pixel.

$$I = (r + g + b)/3 \tag{15}$$

This, however, is a rather coarse approach since it does not take into account the different luminance perception of various colors. The conversion used in this work uses the ITU-R 601-2 luma transform instead.

$$I = 0.299 \cdot r + 0.587 \cdot g + 0.114 \cdot b \tag{16}$$

Figure 20 shows an example of an input image and the results of the two different transforms applied to this image.

Figure 20: Comparison of luminance computations. Input image on the left, conversion after L. Itti et al. (1998) in the center and conversion based on ITU-R 601 on the right.

The color channel in the salience maps of the Itti model is computed based on the concept of double opponency. Center-surround maps are computed for two channels, one red/green channel and one blue/yellow channel which represent the two double opponency channels in the human brain. However, the implementation of the salience map proposed here follows the suggestion of Frintrop (2006). Instead of using two center-surround channels, four color center-surround maps, one for each color, are used. The computation used

to create the basic color feature maps is still as defined by L. Itti et al. (1998).

$$R = r - \frac{g + b}{2} \tag{17}$$

$$G = g - \frac{r + b}{2} \tag{18}$$

$$B = b - \frac{r + g}{2} \tag{19}$$

$$Y = \frac{r + g}{2} - \frac{|r - g|}{2} + b \tag{20}$$

The center surround differences are then computed on six different spatial scales for each color.

$$R(f, c) = |R(f) \ominus R(c)| \tag{21}$$

$$G(f, c) = |G(f) \ominus G(c)| \tag{22}$$

$$B(f, c) = |B(f) \ominus B(c)| \tag{23}$$

$$Y(f, c) = |Y(f) \ominus Y(c)| \tag{24}$$

Where f refers to the fine scale and $c = f + \delta$ to the coarse scale and $f \in \{2, 3, 4\}, \delta \in \{3, 4\}$.

For every spatial scale, the center surround maps are added up across colors yielding one center surround color map for each spatial scale. These maps are downsampled to scale 4 as necessary and added up resulting in the final color conspicuity map. This map is subsequently fused with the intensity and orientation conspicuity maps as defined in section II.C.1.d on page 38.

The original bottom-up salience model uses a normalization scheme which is applied to all center-surround maps before being fused into the conspicuity maps of their respective channel. The same normalization is applied to all conspicuity maps before they are combined into the final salience map (L. Itti et al., 1998). The motivation for normalization is to account for the different dynamic ranges of different modalities and to avoid having locations which are salient in a few maps only suppressed due to noise in other

maps. Different normalization methods were proposed, but none of them are very convincing (Frintrop, 2006; L. Itti et al., 1998; L. Itti & Koch, 2001b). Therefore, an alternate approach is used to take care of the different dynamic ranges. At first, after basic feature extraction, i.e, after creating the intensity map and the four initial color maps, the maps are scaled from 0 to 1 based on the knowledge that the values of the map range from 0 to 255. Then, each time an operation is applied to a map or several maps are fused, the range of the output is determined by considering the possible range of the input maps and the range the resulting maps could have, based on the applied operator. Next, based on this information the intermediate map is scaled to the range of 0 to 1. If, for example, two maps with minimum values of 0 and maximum values of 1 are added to each other, then the values in the resulting map can range from 0 to 2. This resulting map is then scaled to the range of 0 to 1 again by dividing by 2. The scaling does not depend on the actual values in the map, but on the possible minimum and maximum values a map could have based on the operations performed on the input map up to this point. This ensures, that the ranges of all intermediate maps are confined to the range of 0 to 1, and the final salience map will be in the range of 0 to 1 as well. This mechanism not only ensures that all input maps contribute with equal strength, but also that final salience maps can be compared between images. A map with a green dot on a red background, for example, should have a different salience value at the location of the green dot than a red dot on a background with a slightly different shade of red.

The predictive power of the bottom-up salience map derived with this implementation will be compared to the bottom-up salience map derived from the Neuromorphic Vision Toolkit, the actual implementation of the original bottom-up salience model (L. Itti et al., 1998), as well as to the relevance maps. The creation of these will be described in the next section.

2. Relevance Map

The main idea behind the relevance map is to capture information with semantic relevance for the search task. The experiment described in the previous chapter showed, that

scene locations with meaning for the search task attract the eye of observers. Apparently, the meaning associated with the locations is used to inform the search task. Qualitative analysis of the results of the experiment conducted by Wainwright (2008) shows that two types of scene locations receive a substantial amount of fixations. The first locations are the ones at which a ground soldier could take cover, such as small walls and vertical ledges, as well as windows or door frames. The second type of locations would allow a target to blend in well with its environment. This means that the relevance map should capture hiding locations as well as locations at which human ground soldiers would blend in well with the environment.

In order to capture this type of semantically relevant information from the simulation environment, which is the basis for the top-down maps of the proposed eye movement model, two Delta3D-based applications are used. These two applications directly operate with a simulation environment. This environment or geometry is the same as the environment used to create the stimuli displayed to humans in the experiment described in section IV.B on page 117. Both applications, the waypoint explorer application and the intervisibility application, will be described in the following sections.

a. The Waypoint Explorer Application

The main purpose of the waypoint explorer application is to automatically populate a simulation environment, map, or geometry with a large number of waypoints. Usually, the purpose of these waypoints is to guide computer-controlled players in simulations or games. These actors will only go to locations in the environment which are marked by waypoints. Therefore, a dense mesh of waypoints is necessary to help avoid unrealistic behavior of computer-controlled players, if they are supposed to take cover or perform similar sophisticated tasks. For this work, the waypoints are necessary to place target figures at reachable locations in order to compute visibility information pertaining to these targets. This information is subsequently used to create the top-down relevance maps. In order to cover the environment sufficiently well, a dense waypoint mesh is needed in this case as well. This makes the waypoint explorer the ideal application to use. Without an applica-

tion laying out waypoints automatically, a lot of cumbersome manual labor is necessary to acquire reasonably dense meshes. The waypoint explorer allows for the creation of an arbitrarily dense waypoint mesh, requiring only a limited amount of manual work. Although the density of the waypoints can be arbitrarily high, practical considerations usually set an upper bound. The computational power required to derive the intervisibility data increases with the number of waypoints. Therefore, based on the available processing power, runtime considerations and waypoint density need to be reasonably balanced.

The waypoint explorer application is a Delta3D based application working on a scene graph technology based simulation environment. The application is programmed in Python using Delta3D's Python bindings. The application is part of the extras section of Delta3D and available for download at `http://sourceforge.net/projects/delta3d-extras/`.

The waypoint explorer uses an existing waypoint file corresponding to the map that is to be explored. The waypoints defined in the file are used as starting points or seeds for exploration. From each waypoint the explorer tries to go into six different directions, initially from going straight ahead in 60° increments, by a given step size. This results in a hexagonal waypoint mesh. Using the collision model of the Delta3D engine, the application determines whether the explorer can reach the desired location. If this is possible a new waypoint is placed there. Going from the current waypoint to the new waypoint is done incrementally. This ensures that there is a valid path from the current waypoint to the destination, and only then is a waypoint placed at the destination.

The application keeps laying out waypoints as long as there is non-explored space (see Figure 21 for an example). Once the whole space is explored a waypoint file containing all waypoints, the seeds as well as the newly created ones, is written. This file is used by another application, the intervisibility application, to place targets at each waypoint and compute visibility information. This application is explained in the next section.

One idea of the waypoint explorer application is to figure out which areas of the simulation environment could be reached by a human. The stimuli created for this

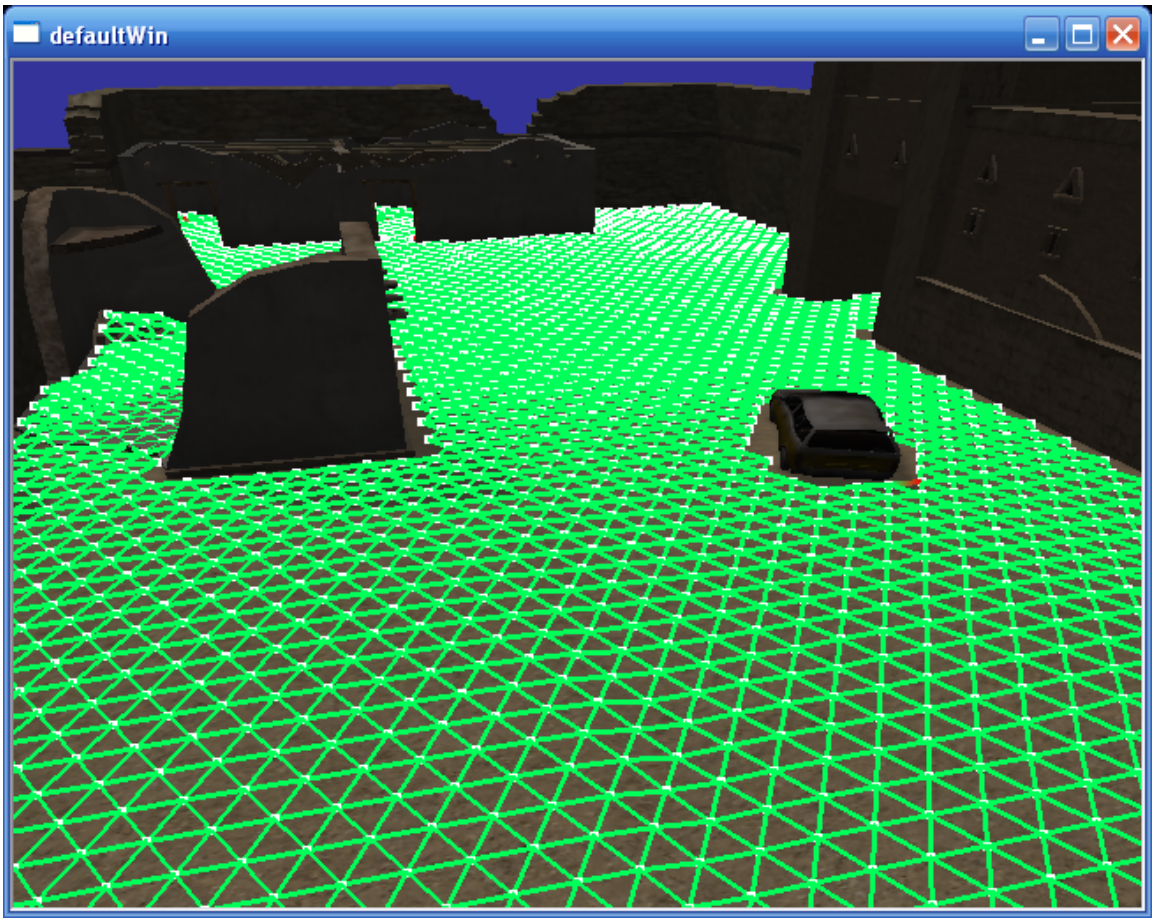

Figure 21: An example of a waypoint mesh laid out in the environment used in this work. White dots are waypoints. The green lines indicate links between waypoints which can be traversed by a person.

work, however, have not been designed with the knowledge of where someone could get to based on the geometry of the simulation environment. Instead, the focus has been on creating scenes which are realistic from an observer's viewpoint. The observer cannot see whether a certain location is accessible, but only whether a target position is realistic or not. This means that by using the stimulus display and design application, targets could be placed at locations which could not be reached without climbing walls, since no stairs are available to get these locations. This would make it impossible for the waypoint explorer to reach these locations from a single waypoint seed placed at ground level. Therefore,

additional waypoint seeds needed to be created on walls, rooftops, and floors otherwise not reachable. In addition to that, a few other cases require manual placement of additional waypoint seeds. First, there are circumstances in which the waypoint explorer cannot find a way through narrow passageways because waypoints are located right in between links of the hexagonal mesh. Then, there are parts of the simulation environment which have no connection to each other. Each of these parts require additional waypoint seeds.

For this work, a total of 101 waypoints needed to be created manually in the used simulation environment in order to cover all locations in the simulation environment at which targets could realistically appear. This seems to be a lot but compared to the total number of waypoints in the waypoint file that was finally used, which amounted to 77751, this number is very small. Only 0.13% of all waypoints had to be placed manually. This number of waypoints is the total of all waypoints the simulation environment was populated with.

b. The Intervisibility Application

The intervisibility application is responsible for deriving the information needed to compute the top-down maps from the simulation environment. In order to compute that information it needs a waypoint file, which has been produced by the waypoint explorer application and it also needs the geometry data which defines the simulation environment. The output of the intervisibility application, the so-called pixelbank, is a three-dimensional data structure containing visibility information of targets placed at the waypoints. In the following paragraphs, the structure of the pixelbank and the visibility data collected is described in detail.

For a given viewpoint resembling an observer's viewpoint, the application renders the scene, which is an image or a frame of a visual simulation. This image shows the simulation environment from the given viewpoint. A scene is rendered several times, and each time a target figure is placed at a different waypoint. One scene is rendered for each waypoint of the waypoint file that is visible from the given viewpoint, with the target figure being placed at this waypoint. Figures 22 and 23 show examples of two scenes

rendered from the same viewpoint. Each scene contains one target, but the targets are placed at different waypoints. The waypoints are not visible in the scenes.

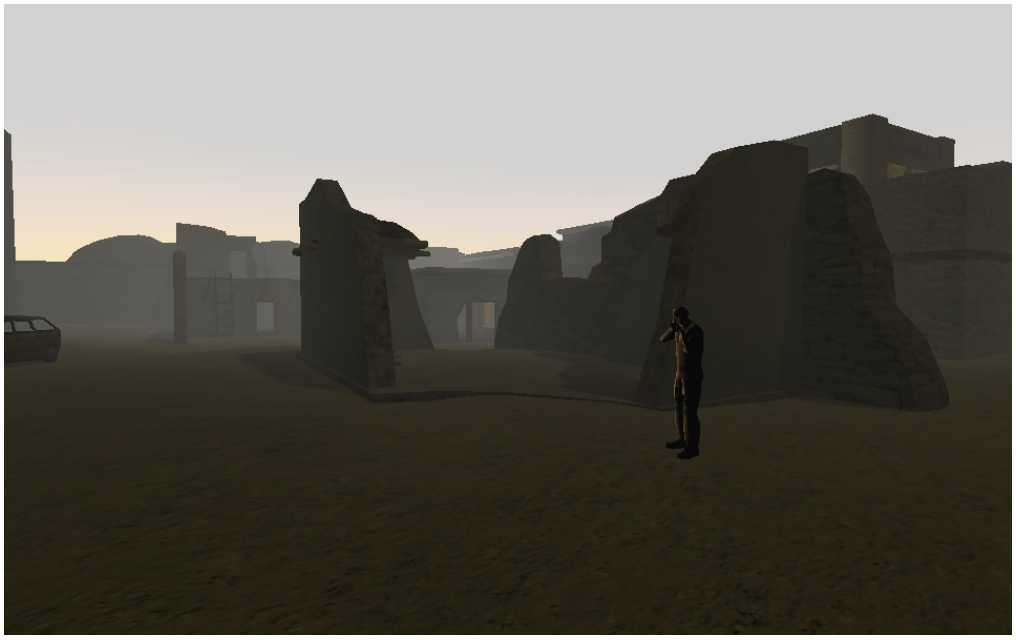

Figure 22: A scene of the environment used in this work rendered with the target at one of the waypoints. The waypoints are not displayed.

Visibility information is collected for that target, and for every pixel of the target, an entry is made at the respective pixel coordinate in the pixelbank. The x- and y-coordinates of the pixelbank are equivalent to the x- and y-coordinates, i.e., the horizontal and the vertical position in the rendered image or frame of that scene. Of course, the waypoints, and with them the targets, have a location in three-dimensional space. Thus, a certain depth in the scene, a z-coordinate representing the distance from the observer's viewpoint is associated with each waypoint and target. Storage of visibility information in the pixelbank needs to account for that. This is done by keeping track of the z-coordinate of each target pixel. The insertion of visibility information for one target pixel is done with respect to its z-coordinate and relative to the z-coordinate of the data already stored in the pixelbank, maintaining order based on the depth information. For one x,y-coordinate,

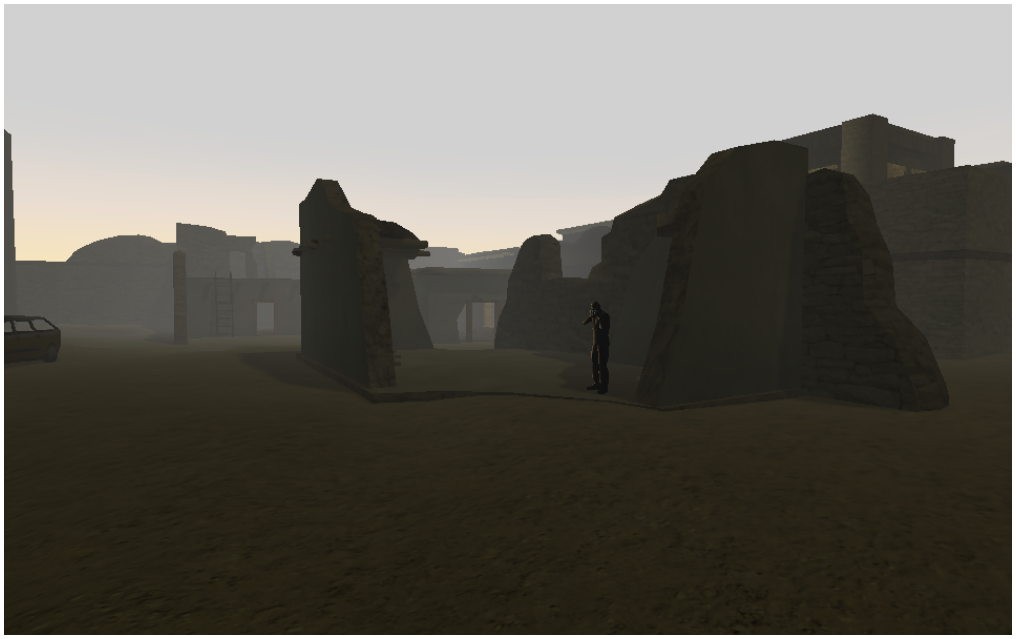

Figure 23: A scene of the environment used in this work rendered with the target at one of the waypoints. The waypoints are not displayed.

larger depth in the pixelbank means the information was derived from a target with larger depth in the scene.

The visibility information that is computed for each target pixel and stored in the pixelbank includes the fraction of visible pixels over the number of total target pixels. This information indicates locations at which a target can hide behind something. If the fraction of visible pixels is zero, no portion of the target is exposed. If it is one, the target is fully exposed. Any number in between indicates, that the target is partially covered. The next information derived is the contrast of the target to its background. High contrasts indicate clearly visible targets and low contrasts indicate targets that blend with the background very well. The contrast computation is performed as defined by C. J. Darken (2007). For each color channel, the target and background 'intensity' is computed using the following formulae:

$$R_T = \frac{1}{n_T} \sum_{p \in T} r^2(p) \tag{25}$$

$$G_T = \frac{1}{n_T} \sum_{p \in T} g^2(p) \tag{26}$$

$$B_T = \frac{1}{n_T} \sum_{p \in T} b^2(p) \tag{27}$$

The background 'intensities' R_B, G_B, and B_B are computed analogously, where the background comprises all pixels within a rectangle around the target, which have a larger scene depth than the target. The rectangle is 5% larger than the smallest rectangle, which would include the target completely.

Then, the contrast is computed for each color channel separately:

$$C_R = \frac{|R_T - R_B|}{R_B} \tag{28}$$

$$C_G = \frac{|G_T - G_B|}{G_B} \tag{29}$$

$$C_B = \frac{|B_T - B_B|}{B_B} \tag{30}$$

and the average of the three contrasts is the resulting contrast value.

$$C = \frac{C_R + C_G + C_B}{3} \tag{31}$$

In addition, the intervisibility application captures information such as the number of visible target pixels and target detection probability according to the ACQUIRE implementation of C. Darken and Jones (2007). Currently, none of this information is used for the purpose of this work.

c. Generation of the Relevance Map

From the pixelbank two top-down maps are computed. One map, which is based on the fraction of visible pixels, contains the information about hiding locations. The second map, based on the contrast information, indicates locations at which targets blend in well with the environment.

The hiding location map is derived from the pixelbank by taking the minimum fraction of visible pixels from the list at every pixel. This yields a two-dimensional map ranging from 0 to 1. The width and height of this map are the same as the width and the height of the image rendered from the simulation environment that correspond to the respective scene. Pixels with small numbers indicate locations at which targets are occluded and therefore likely hiding locations (see Figure 24). This map is inverted, mapping the range of 0 to 1 to the range of 1 to 0 such that 0 represents a fully exposed target and the numbers close to 1 indicate hiding locations (see Figure 25).

Figure 24: The predecessor of the hiding location map as derived from the pixelbank. Good hiding locations are indicated by black pixels and locations at which targets are fully exposed are white.

Figure 25: The hiding location map of one scene. White pixels indicate likely hiding locations.

Similarly, the contrast map is a two-dimensional map with the same width and height as the hiding location map and the pixelbank. For each pixel, the minimum contrast is picked from the list at this pixel. The range of pixel values of this map starts at 0 and can be arbitrarily high. In practice, however, the numbers range from 0 to 1 in most of the cases (see Figure 26). Therefore, all values above 1 are set to one and the result is mapped to the range of 1 to 0, 0 meaning very high contrast and 1 no contrast to the background at all. In practice, the latter was not observed for the scenes used in this work. Thus, high numbers represent locations at which the target can blend in well with the environment and 0 represent locations at which a target stands out well from the background (see Figure 27).

The final relevance map is derived by additively combining the hiding location map and the contrast map. Figure 28 shows an example of a relevance map and Figure 29 illustrates the derivation of the relevance map from the pixelbank. The relevance maps will be assessed as to how well they predict fixations by comparing them with eye-tracking

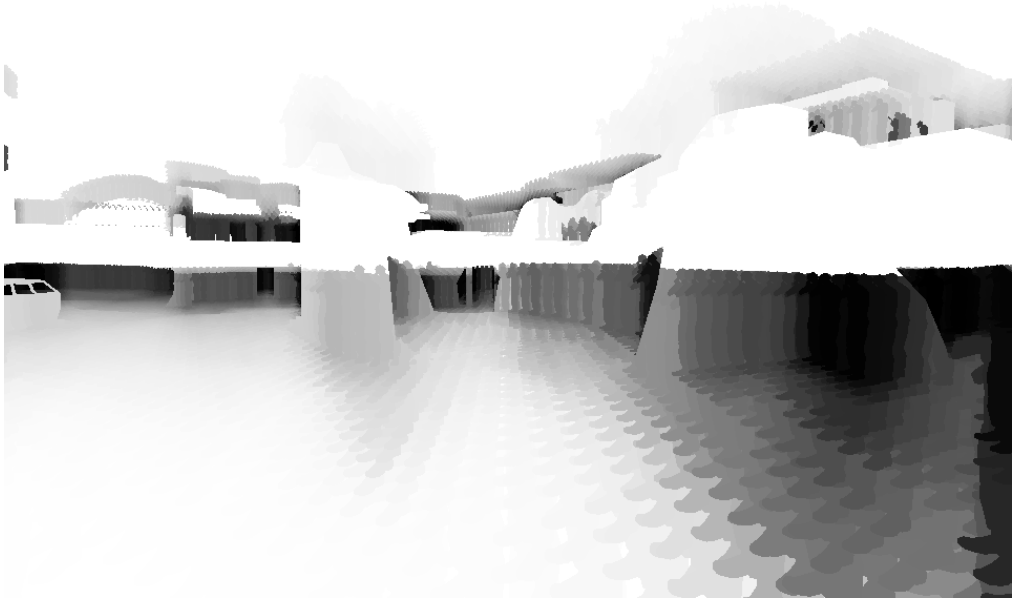

Figure 26: The predecessor of the contrast map as derived from the pixelbank. Locations at which targets blend in well with the background are indicated by black pixels and locations at which targets have a large contrast to the background are white.

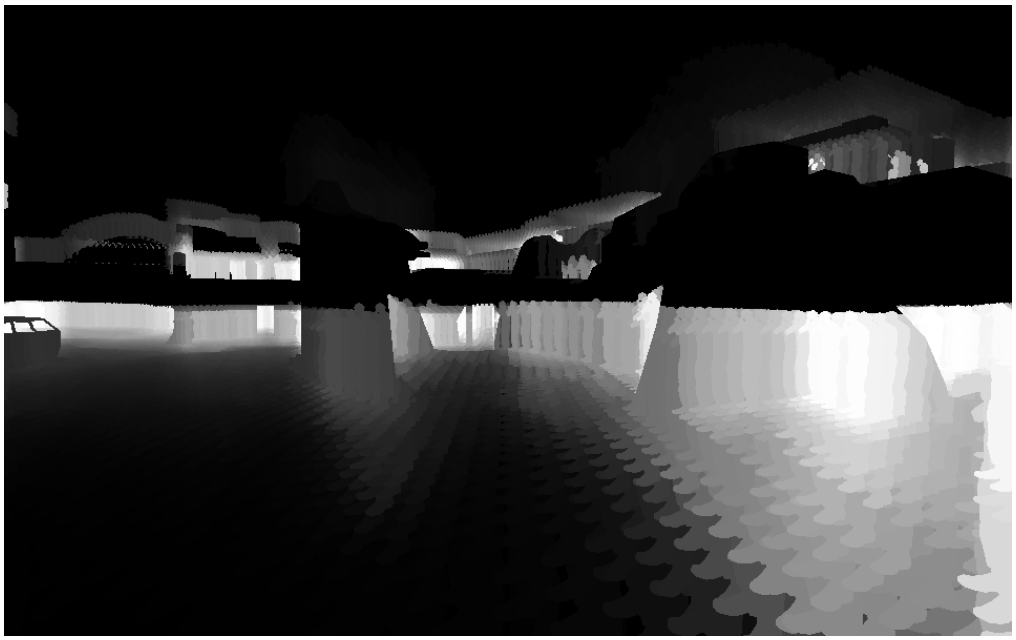

Figure 27: The contrast map of one scene. White pixels indicate locations at which a target blends in well with the background.

data. The data was collected from participants viewing realistic scenes containing one to four targets. These scenes were used to derive the relevance maps. In the next section, this search and eye-tracking experiment is described in detail.

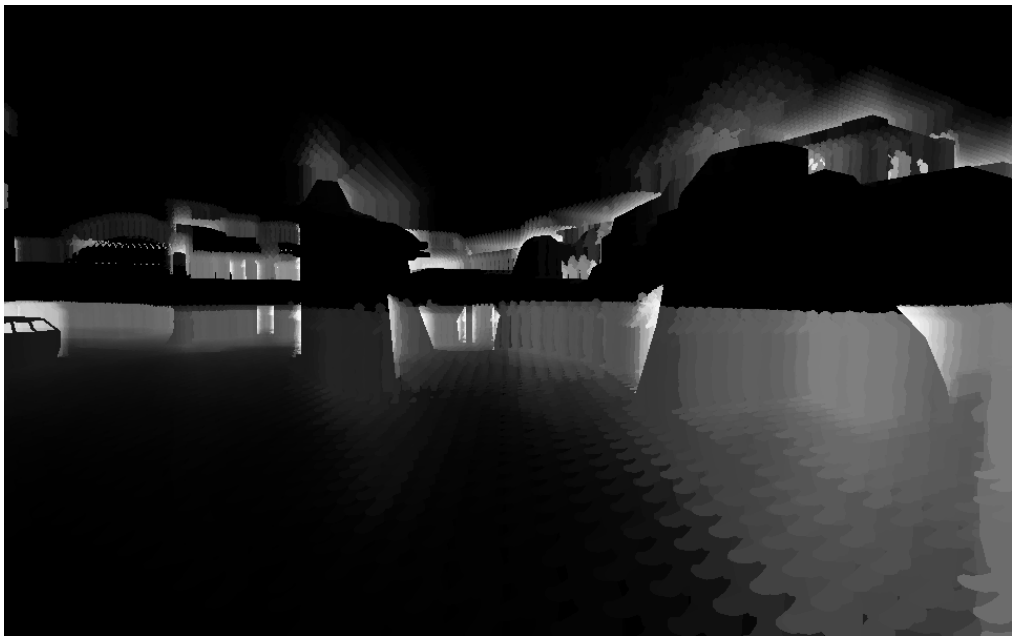

Figure 28: The relevance map for one scene. White pixels indicate the relevant scene locations.

B. EYE MOVEMENT EXPERIMENT IN NATURALISTIC SCENES

1. Participants and Apparatus

The participants and the apparatus were the same as described in section III.B.

2. Stimuli

The stimuli presented in this experiment were designed using a stimulus generation and display application developed at the Naval Postgraduate School based on the Delta3D game engine. In contrast to the previously used stimuli (section III.B.2), the ones for this experiment were more realistic. They were designed as scenes a ground soldier could possibly encounter in an urban environment. The targets in the scenes were enemiy soldiers

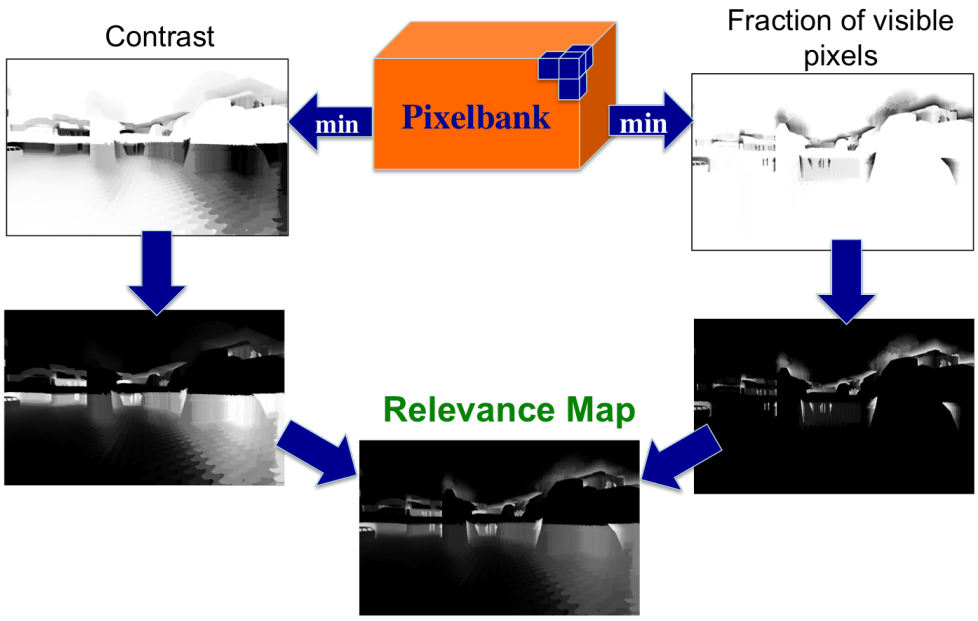

Figure 29: Derivation of the relevance map from the pixelbank.

in camouflage uniform hiding in structures, behind walls, or other objects in the scene. Enemy soldiers could also be present in open areas. Each scene contained one to four targets. The targets used were the same as in the previous experiment, but they could appear in four different postures: standing, kneeling, crouching or prone. Although more realistic, the scenes were static images without any movement. A total of sixteen scenes were presented for a maximum of fifteen seconds each. Figures 30 and 31 show two of the sixteen stimuli.

3. Design and Procedure

After the completion of the experiment described in the previous section, participants continued with this experiment. They were briefed that they would view more realistic scenes containing one to six instances of the familiar target in the following part of the experiment. Participants were also informed that the targets could appear in the open or that they could be hiding or taking cover behind other objects, and that the targets could assume four different postures. In order to familiarize the participants with the possible target

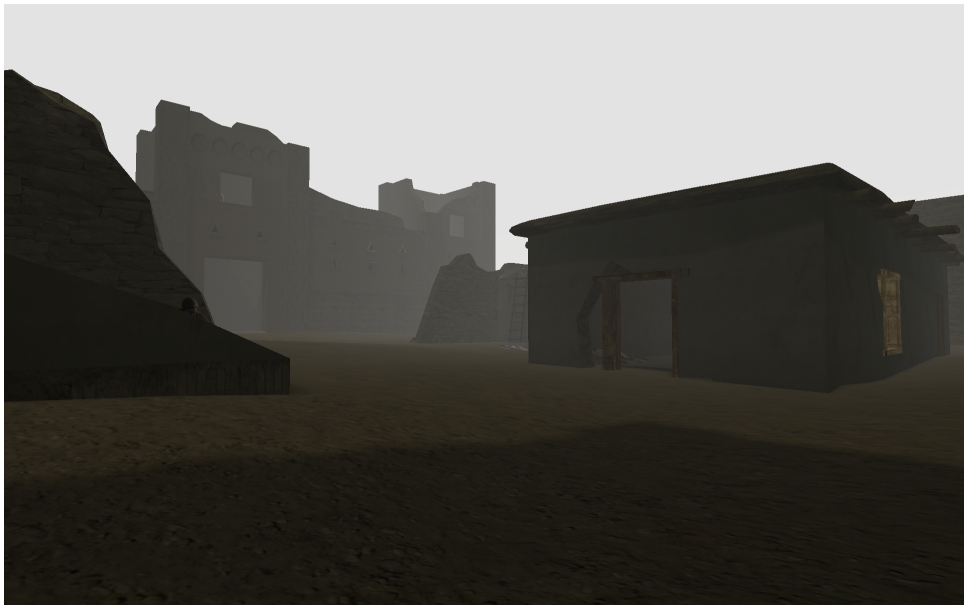

Figure 30: Example stimulus with four targets.

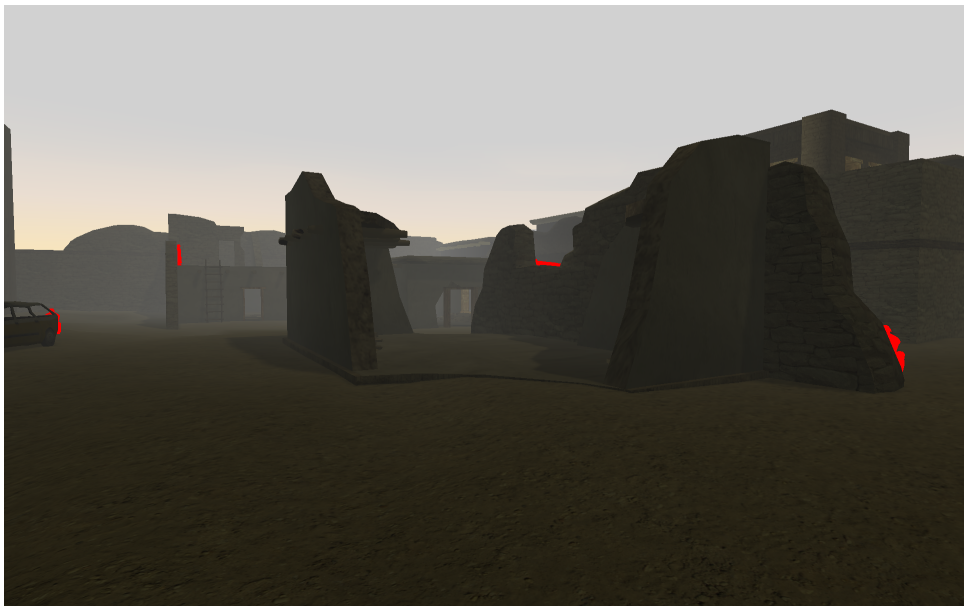

Figure 31: Example stimulus with four targets. In order to highlight the target locations, the targets are false color rendered. Stimuli were not presented to participants in this way.

appearances, examples of the different postures as well as examples of partially occluded targets were presented. Then, the participants viewed one training scene before starting with the experiment. The scenes were displayed until participants indicated that they had found all targets by saying 'next', but not longer than 15 seconds. Before each scene, the same fixation cue as in the previous experiment was presented to participants. They were asked to fixate the cue until the search scene was displayed.

Although a maximum of four targets were present in each scene, participants were told that there could be one to six targets in order to avoid search termination based on the number of targets found. Also, the instructions stressed that it was important to find all targets by pointing out that missed targets could be of continuous danger in future.

Before the start of the experiment, the participant's understanding of the task was tested by asking a few questions addressing the key points of the task. After that, the sixteen scenes were presented without any interruption.

4. Fixation Determination

The fixation determination is performed by first finding saccade starting points and end points. Then, all gaze points in between saccades are considered part of one fixation. The fixation location is established by computing the center of gravity of all gaze locations belonging to the fixation. The detection of saccade start and end times is performed using a speed threshold of 8.75° of visual angle per second over two consecutive gaze points. Visual inspection of scene overlays shows that this threshold separates saccades from fixations sufficiently well. It is not necessary to employ the direction change mechanism as described in the previous chapter.

C. RESULTS AND DISCUSSION

The quality of the recorded eye-data was rather mixed for the search experiment in realistic scenes. This quality varied across subjects and also within subjects. For some scene/subject combinations the quality of the eye-tracking was really poor, and therefore

they needed to be excluded from the analysis. Every scene/subject trial was examined visually and the eye-tracking quality was judged. If it was noticed that the eye-tracking quality was not good enough, the individual trial was excluded from the analysis. The judgement was based on the fixations close to the locations participants clicked on to indicate target presence. If fixations on these clicks had been further than 2° of visual angle from the click location, the scene/subject combination was excluded. This was based on the eye-tracking error of around 1° of visual angle plus some possible offset of fixation from the actual target location still achieving sufficient resolution for the participant, even if the central part of the fovea was not directly placed at the click location. This decision was based on the assumption that clicking on targets is not performed using peripheral vision.

1. Fixation Maps

In order to compare the fixations with the salience and relevance maps, fixations on one scene over all participants are fused into one fixation map per scene. The fixation maps have the same width and height as the stimuli presented: 1920x1200 pixels. The fixation maps are binary maps containing either values of 0 or 1. Each location of the fixation map for which a fixation was recorded is set to 1. All other pixels of the fixation map are set to 0. This means that a 1 in the fixation map indicates a fixated location and a 0 indicates a location which was never fixated. Figures 32 and 33 show the fixation maps as heatmaps overlaid on a scene and a relevance map, respectively. The heatmaps are derived by blurring the fixation maps with a 49×49 pixel large Gaussian kernel. The size matches the experienced eye-tracking error of about 1° of visual angle.

2. Comparison Metric

The fixation maps are compared to the salience and relevance maps using the area under the curve (AUC) of a receiver operating characteristic (ROC) curve following Tatler, Baddeley, and Glichrist (2005) and Einhäuser, Spain, and Perona (2008). An ROC curve plots the false positive rate by the hit rate of a classifier or predictor for varying thresholds applied to that classifier. The hit rate is also referred to as the true positive rate.

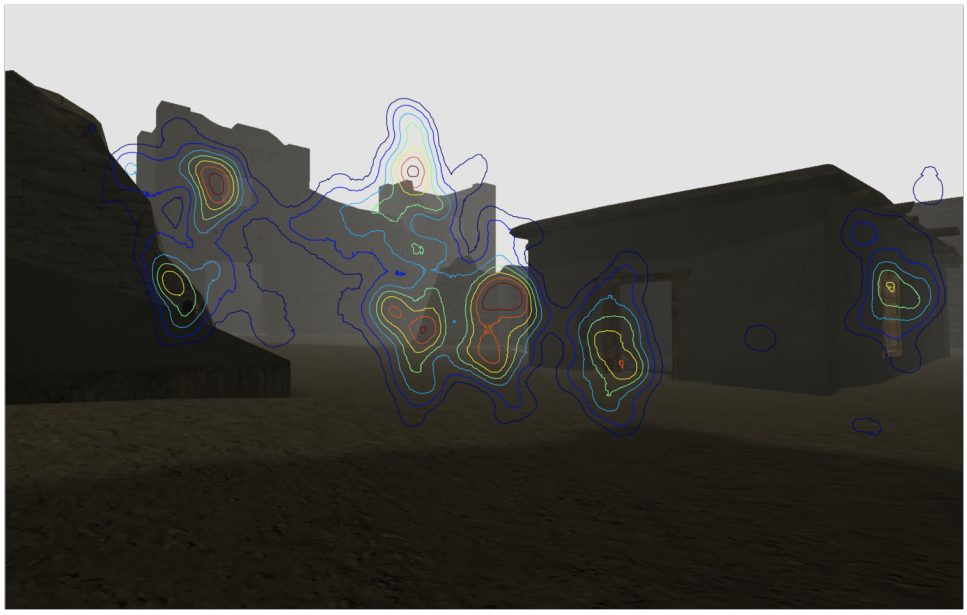

Figure 32: A fixation heatmap, indicating the fixation density on one scene over all participants, superimposed on a stimulus.

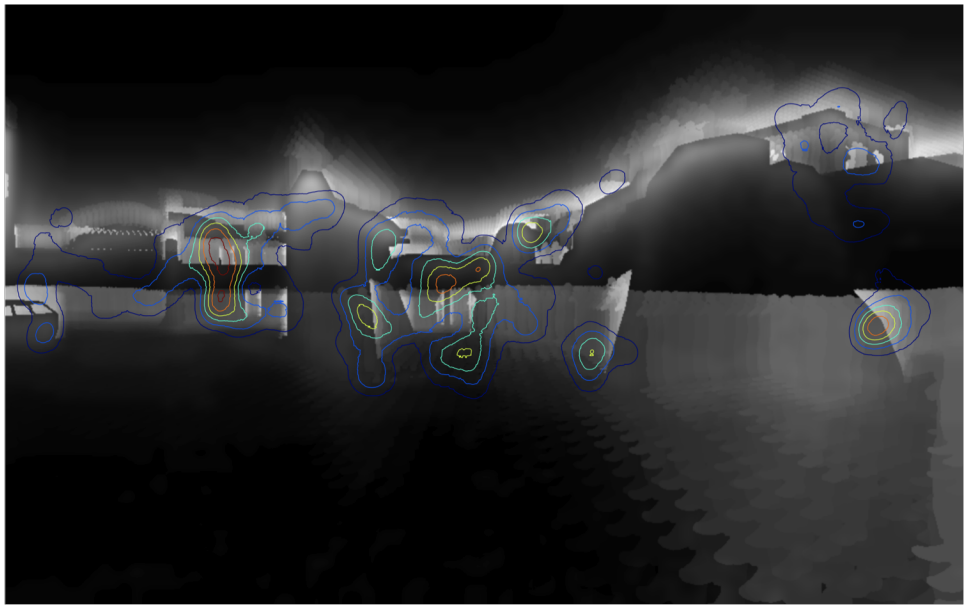

Figure 33: A fixation heatmap, indicating the fixation density on one scene over all participants, superimposed on a relevance map.

$$\text{false positive rate} = \frac{\text{negative instances classified as positives}}{\text{total number of negative instances}} \tag{32}$$

$$\text{hit rate} = \frac{\text{positive instances classified as positives}}{\text{total number of positive instances}} \tag{33}$$

For the fixation maps, the total number of negative instances for one scene are the number of zeros in the fixation maps, which are all the locations that were not fixated by any participant. Conversely, the total number of positive instances for one scene is the number of ones in the fixation map. These are all the locations that were fixated by at least one participant.

The salience maps and the relevance map are treated as predictors of fixations. All values in the map above a certain threshold indicate that this location will be fixated. All values below that threshold indicate that these locations will not be fixated. The locations which are above that threshold and are marked as fixations in the fixation map are hits based on that threshold. All locations which are above the threshold and not marked as fixations in the fixation map are false positives. This assumption, however, is very conservative, since in reality a fixation covers more than just one pixel. Pixels with values above the threshold that are not fixated but lie in the immediate vicinity of the fixation location, will be counted as false positives and not as hits. As a result, the values of the metric used (area under the ROC curve, described in the following paragraphs) will be lower than they should be. However, the proposed comparison metric is still appropriate, since the evaluation of the maps is based on a comparison of the values, not their magnitudes.

Based on the numbers of hits and false positives the false positive and hit rate for this threshold can be determined and they establish one point of the ROC curve. Varying the threshold over the range of the predictor, in this case the salience and the relevance maps (ranging from 0 to 1), yields a set of points forming the ROC curve. A more detailed explanation of the ROC curve creation can be found in Fawcett (2006). Figure 34 shows the ROC curves for two of the sixteen scenes. Four curves are graphed, one for each of the

predictor maps being compared. Figure 35 shows a plot of all ROC curves of the sixteen scenes and the four predictor maps.

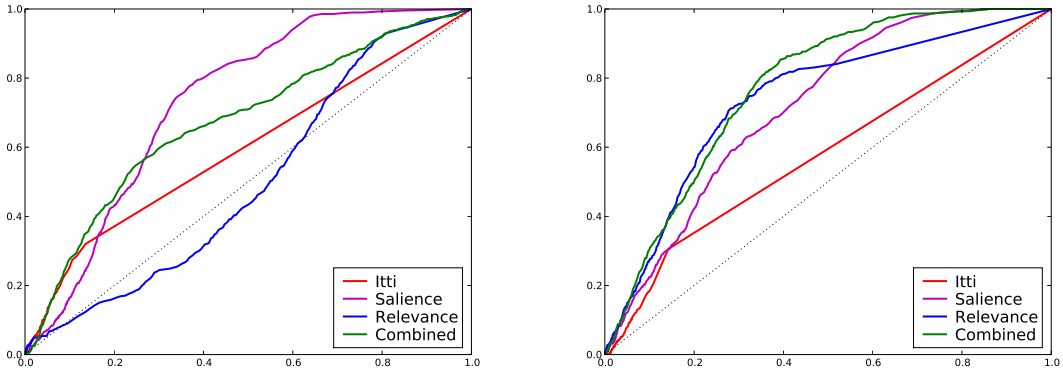

Figure 34: ROC curves of the four predictor maps of two of the sixteen scenes.

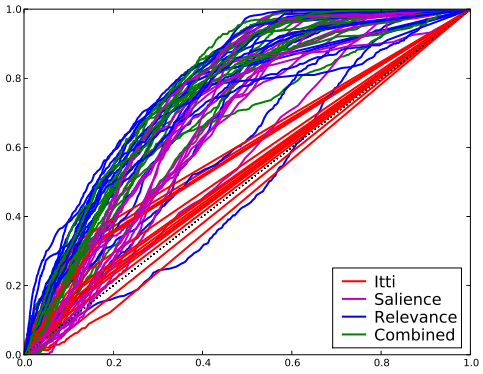

Figure 35: ROC curves of all sixteen scenes and all four predictor maps in one image. It can be clearly seen how the relevance map and the map combining relevance and salience dominate the pure salience maps.

One way of employing the ROC curve to compare classifiers or predictors is to use the area under the curve (AUC). The AUC is a convenient scalar value that easily shows which of two AUCs is larger and which is smaller. The important thing is, however, that the AUC has a very interesting statistical property. It is equivalent to a Wilcoxcon rank-sum test. This means that the AUC represents the probability with which positive instances can be distinguished from negative instances (Hanley & McNeil, 1982). Applied to the

salience and relevance maps, this means that the AUC tells how well these maps correctly distinguish between fixations and non-fixations. Therefore, it is suitable for comparing the predictive power of the bottom-up and top-down maps.

For the comparison of the fixation maps with the predictor maps the eye-tracking error needs to be taken into account. Since the fixation maps are intended to be binary maps, it is not appropriate to apply correction mechanisms to these maps. Instead, the predictor maps are convolved with a Gaussian kernel with the size of 1° of visual angle, which is approximately the size of the eye-tracking error. This results in smoothed maps which contain information of the surrounding pixels within 1° of visual angle at each pixel location.

3. Results

A total of four maps are compared to the fixation maps of each scene. This yields one AUC per map and per scene, i.e., 16 AUCs for each map. The assessed maps are the bottom-up salience map of the original implementation of the Itti model[1] (referred to as the Itti map from here on); the re-implemented bottom-up salience map, which follows the specification of the Itti model as described in section IV.A.1 on page 103, the relevance map and an additive combination of the re-implemented bottom-up salience map and the relevance map called the combined map. This combined bottom-up/top-down map is computed by adding up the two input maps both weighted with 0.5.

In order to be a useful predictor, the AUC of the maps needs to be larger than 0.5. An area of 0.5 would be achieved by random guessing. The average areas under the curve of the Itti map (μ=0.54, σ=0.04, p=0.0007), the salience map (μ=0.69, σ=0.05, p<0.0001), the relevance map (μ=0.72, σ=0.07, p<0.0001) and the combined map (μ=0.74, σ=0.03, p<0.0001) all statistically significantly exceed 0.5 (see Table 2). This means that all of them predict eye fixations better than chance. However, it is apparent that there is a large

[1]Derived from `http://ilab.usc.edu/toolkit/downloads.shtml`, last accessed 17:17 22APR2009

difference between the AUCs of the four maps. Therefore, the maps are compared to each other in order to see if they differ in their predictive power.

Predictor Map	avg. AUC
Itti	0.54 ($\pm$0.04)
Salience	0.69 ($\pm$0.05)
Relevance	0.72 ($\pm$0.07)
Salience + Relevance	0.74 ($\pm$0.03)

Table 2: Average area under the ROC curve (AUC) of the four predictor maps.

Two different methods are used for the map comparisons. The first method is a t-test of the means of the area under the ROC curve. The second method compares the AUC of one scene between maps and counts how often one map performs better than the other. If one map is doing as good as another, the ratio of cases in which it is doing better would be 0.5. If this ratio is different from 0.5, one map is doing better than the other. Whether there is a significant difference from 0.5 is assessed using a sign test based on a significance level of 0.05.

Based on a t-test of average AUC, the Itti map is doing significantly worse than the salience map ($p<0.0001$), the relevance map ($p<0.0001$), and the combined salience and relevance map ($p<0.0001$). This means that based on the average AUC the Itti map is the worst predictor of eye fixations. The average AUC of the salience map is statistically not significantly different from the average AUC of the relevance map ($p=0.0930$). However, the p-value is still rather low and with 0.72 the average AUC of the relevance map is higher than the average AUC of the salience map which amounts to 0.69. This indicates a trend of the relevance map being a somewhat better predictor than the salience map. Comparing the salience map with the combined salience and relevance map shows that the combined map is doing significantly better than just the salience map ($p=0.0052$). Very interestingly, there is no significant difference between the average AUCs of the relevance map and the combined salience and relevance map ($p=0.2386$). Together with the fact that the combined

map is a better predictor than the salience map, this is further indication that the top-down map is doing better than the bottom-up map in predicting eye fixations.

The next assessment compares all pairs of maps on a per scene basis, counting how often each of the maps has a higher AUC, i.e, how often each map is predicting eye fixations better for a certain scene. The comparisons are based on a sign test using a significance level of 0.05. Comparing the Itti map with the salience map shows that the Itti map is doing better in no scene, and the salience map is doing better in all 16 scenes. The same result is found for the comparison of the Itti map with the combined relevance and salience map. This difference is statistically significant ($p<0.0001$). As compared to the relevance map, the Itti map is doing better in 1 case and the relevance map in 15 cases. Again, the difference is statistically significant ($p=0.0003$). Clearly, the Itti map is inferior to all other maps. Looking at the salience map, one can see that it predicts eye fixations better than the relevance map on 4 scenes, whereas the relevance map is a better predictor for 12 of the total 16 scenes. A sign test of this ratio shows statistical significance ($p=0.0262$). The salience map is also a worse predictor than the combined relevance and salience map. The proportion here is 1:15, which is significant as well ($p=0.0003$). This means that the salience map performs better than the Itti map only. The other two maps, which both contain information about semantically relevant scene locations, are better predictors of eye fixations than the salience map. Finally, the comparison of the relevance map with the combined map shows that each map is doing better than the other for 8 of the 16 scenes. This proportion is obviously not showing a difference of predictive power ($p=0.5$). A summary of these results can be found in Table 3.

4. Discussion

The most apparent result of the map comparisons is that the Itti map, which is the most well-known model of visual attention allocation and eye movements, is outranked by all other maps. This bears the question, whether the used stimuli are special in a certain way and not representative of actual environments such that the Itti map is doing worse than it would on real world stimuli. Previous research of eye movements on real world

	Itti	Salience	Relevance	Salience + Relevance
Itti		0*	1*	0*
Salience	16*		4*	1*
Relevance	15*	12*		8
Salience + Relevance	16*	15*	8	

Table 3: Comparison of the prediction performance of all maps with all other maps. Each number indicates for how many scenes the area under the ROC curve (AUC) was larger for the map of the row as compared to the map of the column. Asterisks indicate statistical significant difference based on a sign test (significance level α=0.05).

photographs using the AUC as a metric as well obtained very similar results (Einhäuser, Spain, & Perona, 2008). They report that the Itti map predicts fixations above chance (AUC > 0.5) in 77 out of 93 scenes, which is 82.8% and an average AUC of 57.8% $\pm$ 7.6%. For the scenes in this experiment, the Itti maps predict fixations above chance in 87.5% of all scenes (14 of 16), and the average AUC amounts to 54.0% $\pm$ 4.1%. This means that the performance of the Itti maps in the experiment of Einhäuser, Spain, and Perona is almost exactly the same as the performance observed here. Since these results were derived from two completely independent experiments this has two implications. First, it means that the stimuli presented in this experiment are equivalent to real world photographs with respect to the general eye movement patterns they elicit and can therefore be considered representative of real world scenes. Second, the results provide an estimation of the general performance one can expect from this particular implementation of the Itti model.

Einhäuser, Spain, and Perona (2008) present work that improves the Itti model by looking at the salience of objects. Due to this alteration the average AUC increases to 65.1% $\pm$ 10.6%, and is in the range of the AUC of the re-implementation of the salience maps (68.9% $\pm$ 4.8%). This observation is very interesting because it shows that changing some implementation details of the original model results in a statistically significant better performance. Performance which is even slightly better than the conceptual improvements

to the original Itti-model of Einhäuser, Spain, and Perona, which is still based upon the original implementation of the model.

Another strong improvement can be observed for the predictive power of the relevance map. The average AUC of the relevance map ($71.9\% \pm 7.1\%$) is larger than the average AUC of the salience map ($68.9\% \pm 4.8\%$), and the relevance map outranks the salience map on a statistically significant number of scenes. This shows very clearly that semantically relevant scene locations are better predictors of eye fixations than visual salience, which is consistent with the results found in the search experiment described in the previous chapter. In addition to that, the result shows that the novel approach of using information from the simulation environment to determine the semantically relevant locations is highly effective.

An even better predictor than the relevance map alone is the combined salience and relevance map. This map outperforms the salience map on 15 scenes and reaches an average AUC of $74.1\% \pm 3.0\%$. This is the expected result based on the experiment described in the previous chapter which showed that both visually salient distractors as well as task-dependent influences affect the eye movements. It is interesting that the combined map does not perform statistically significantly better than the relevance map although the average AUC of the combined map is higher than the average AUC of the relevance map. Also, the fact that the combined map shows an improvement over the relevance map when compared to the salience map would make one think that the combined map would fare better than the map containing the semantically relevant information alone.

Looking at the individual scenes more closely reveals that for scenes in which one of the constituent maps has poor performance, the combined map will perform worse than the best constituent map. In cases in which the performance of both maps is rather good, the combined performance increases. Since the salience map is doing worse than the relevance map for most of the scenes, the salience map can reduce the performance of the combined map as compared to the relevance map alone. In contrast, the contribution of the top-down

map to the salience map in the combined map improves performance as compared to the salience map alone.

In other words, there are scenes for which the bottom-up information is the governing factor. In this case the salience map predicts fixations better than any of the other two maps. This was the case for only one of the 16 scenes assessed in this work. Fixation heatmaps overlaid on the relevance map and the salience map of this scene are shown in Figures 36 and 37, respectively. In Figure 36 it can be seen that there is a large area in the foreground at which the target apparently had low contrast. This seems to be the reason that makes the relevance map a bad predictor of eye fixations for this scene. Then, there are scenes for which the task influence is the governing factor and the relevance map is the best predictor. Lastly, there are scenes, where both bottom-up and top-down information play a significant role, which yields better performance of the combined map than any of the individual maps. The results indicate that in the minority of the scenes, the bottom-up information is the governing factor. In this experiment, there is only 1 scene for which the visual information governs the eye movement, 8 scenes in which the semantically relevant information takes precedence, and 7 scenes for which an equal combination of the two yields best results. This highlights the importance of the semantically relevant scene location over visually salient locations.

This asymmetric behavior of the salience map, relevance map, and combined map is in accordance with the observations of the experiment described in the previous chapter. Although the task demands played the most important role for the eye movement allocation, the salient distractor could draw the eyes at any time. The fixations of the participants looking for targets in the realistic scenes in the second experiment show that the relevance maps clearly have stronger predictive capabilities than the salience maps. On the other hand, visual salience can take precedence, as could be observed for one of the scenes for which the salience map clearly outranked the relevance map. It can also be the case that in one scene the visual salience draws the attention as strongly as the semantically relevant

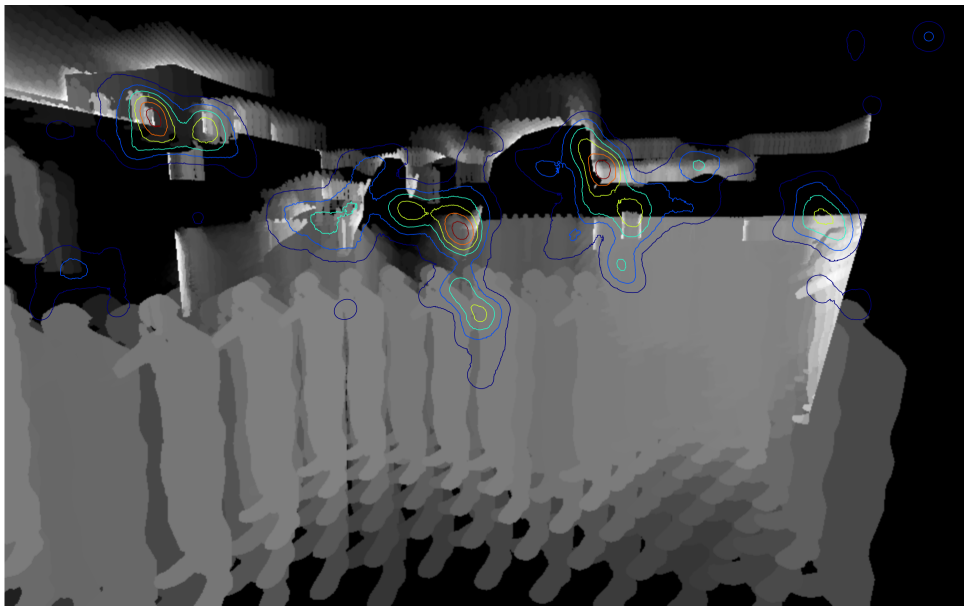

Figure 36: A fixation heatmap of the fixations on the scene for which the relevance map had its worst prediction performance. The heatmap is superimposed on the relevance map.

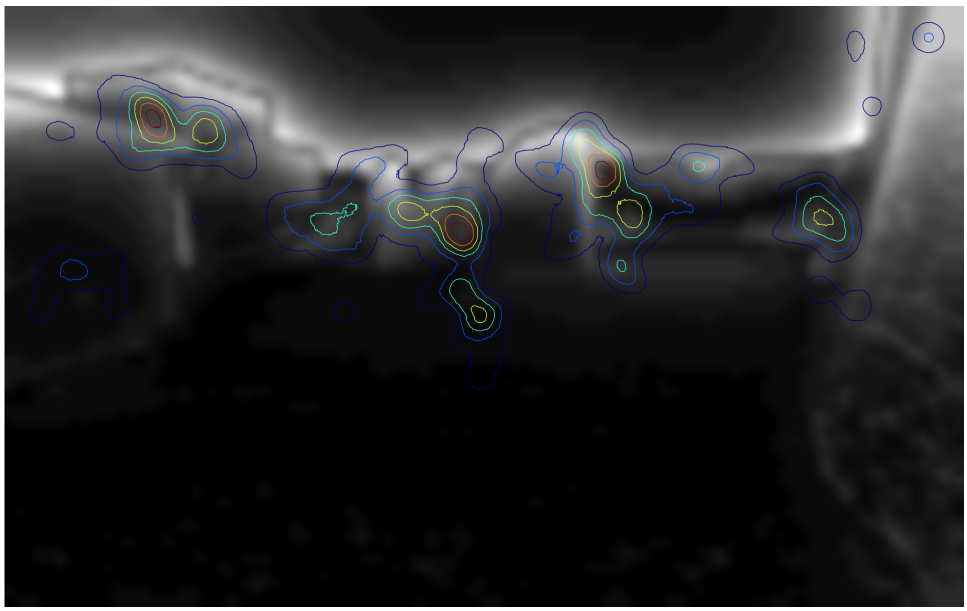

Figure 37: A fixation heatmap of the fixations on the scene for which the relevance map had its worst prediction performance. The heatmap is superimposed on the salience map, which had good prediction performance for this scene.

scene locations do. In this case the combined map shows better performance of eye fixation prediction than the salience map or the relevance map on their own.

Similar to these findings, the contextual guidance model shows that scene features co-occurring with the target can be learned, and these features can be used to modulate a salience map (Torralba et al., 2006). The major difference between this model and the work presented here is that the contextual guidance model has to rely on low-level visual features and can only capture co-occurrence of targets with visual scene features. In contrast, the relevance maps introduced in this work, which capture the meaning of scene locations for the search task, predict eye fixations very well on their own. In addition to that, the contextual guidance model needs to be extensively trained on scenes for which the target locations need to be marked manually, such that the model can learn the association between targets and low-level scene features.

Furthermore, how well the model predicts fixations was assessed by looking at the first 5 fixations only since their "models are expected to be more representative of the early stages of the search, before decision factors start playing a dominant role" (Torralba et al., 2006, p. 778). The model presented in this work, on the other hand, considers all fixations over a rather long time span of 15 seconds, and the results show that it is very well capable of predicting fixation locations for the entire duration. Apparently, capturing the scene locations that are meaningful for the search task addresses the "decision factors" the contextual guidance model is lacking.

It is important to note that the relevance maps do not capture information about target features. Since it was shown that image locations which contain target features receive a higher proportion of eye fixations than locations which do not (Pomplun, 2006), one can expect to get even better results if this top-down modulated bottom-up information would be included in the model. Essentially, this means that there are three factors which should be considered relevant for eye fixation allocation:

1. Bottom-up information which are the visually salient scene regions.

2. Top-down modulated information, i.e., the scene locations containing target features.

3. High level task influence; the scene locations with semantic relevance for the current task.

D. CONCLUSION

In summary, it becomes evident from this research effort that the most influential factor for the prediction of eye fixations is the set of semantically relevant scene locations. In addition, this work presents a novel method which allows the direct extraction of semantically relevant information from a simulation environment. This information is fused into the relevance map, which has very good prediction performance.

Another major finding is that these semantically relevant locations predict eye fixations very well on their own, with even better performance if the relevance map is combined with the salience map.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSIONS AND FUTURE WORK

A. CONCLUSIONS

The results of the first experiment clearly show that in the absence of a distractor the target is fixated right after the first saccade, even at high eccentricity and very low saliency. This is an important baseline because it shows the performance that can be expected from the human visual system in the best case. In the presence of the distractor it becomes apparent that the top-down signals of the target have priority over the bottom-up influences. However, the distractor has still the capability of drawing the attention at any time, and the search performance is reduced when a distractor is present.

In addition, it is apparent that the properties of the target and distractor hugely influence the search performance and eye movements. This could be observed for a variety of metrics. A similar observation was made for initial saccade latency (Born & Kerzel, 2008). This, however, was the only metric this effect could not be observed for. It has already been pointed out by Born and Kerzel that the effect they observed might change due to task demands. It is very important to note that the research presented here, in which there was a clear task demand with very specific instructions, showed a very similar effect to the one reported by Born and Kerzel and thus significantly extends the previous findings.

This novel finding has tremendous implications for the creation and usage of saliency maps. It cannot just be assumed that the maximum values of the saliency map receive the fixations. In addition, any normalization scheme that promotes only the strong peaks is most likely misguided. This type of normalization scheme has been used by various saliency-based visual attention models (Frintrop, 2006; L. Itti et al., 1998).

In the past, these influences of target and distractor properties have not been examined very well and they have not been included into visual attention models. It is now clear that visual attention models cannot only use saliency maps which highlight the most salient locations of a scene, but that the whole spectrum of saliency values needs to be kept

in the saliency map. Taking the experiment results into account, the correct way of using saliency-based bottom-up information is by applying a non-linear transformation to the currently used saliency maps in order to have the spots with intermediate saliency attract the gaze more than the lowest and the highest spot. The difficulty is that the gained results do not provide enough information which would directly allow the transformation for the saliency maps to be derived. This would need to be examined more closely in future.

Since the eye movements and search performance also depend on the target properties, these need to be taken into account as well. First of all, the distance of the fixation location to the target location plays a role; larger distance requires more time and more fixations to reach the target. The opposite is true for the saliency. Therefore, this information needs to be included if a model not only predicts fixation sites, but also fixation order and fixation duration. This also needs to be examined more closely. More details about future work dealing with modeling fixation order will follow later in this chapter.

The results of the experiment also show that the hiding location influences eye movements in a way that is considerably different from the influence of the visually salient distractor. First of all, the hiding location interacts with the target saliency. If the target saliency is above a certain level, the hiding location eccentricity does not have an effect on search performance. If, however, the target saliency is very small, higher hiding location eccentricity reduces search performance. This clearly indicates that participants try to speed up the search process by looking for the target at a location which is indicative of target presence if the target cannot be spotted right away. This location is a part of the scene which is meaningful for the search task. In contrast to the effect of hiding location eccentricity, the distracting effect of the visual salient distractor is reduced with higher eccentricity. Since the influence of the visually salient distractor is reflexive in nature, the influence of the hiding location cannot be.

This is the most important result of the experiment and it shows that semantically relevant scene information can be extracted, processed, and used to guide eye movements and inform the search process. This behavior does not need to be trained, that is partic-

ipants do not have to explicitly learn this association during the experiment. Rather, this information seems to be conceptually stored and immediately available during the search.

This is very interesting and bears the question if it is possible to establish these associations for complex tasks through training. If so, this could be employed for the training of ground soldiers to more quickly spot and examine locations which are important for finding a target or any other important task. Especially, this might be tremendously helpful for IED prevention.

The experiment findings, which show that bottom-up signals, top-down signals pertaining to the target and top-down signals associated with meaningful scene locations are important aspects, also seem to indicate that there is some ranking of these three contributors. It has been shown that the hiding location attracts the gaze with a much higher frequency when the target is barely visible. This indicates that the semantically relevant scene locations are examined only if the target cannot be spotted right away. If the target is well visible, it is rare that it is not fixated in the first fixation, or that the first saccade is actually directed towards the target. The visually salient distractor can capture the attention at any time but overall it does not show general overriding capability. It seems that target presence has the largest influence, then the semantically relevant location, and finally the visual information. However, as already indicated, this does not seem to be a strict ranking since the visually salient distractor could draw the attention any time. To sort out the details of how these interactions work, more research needs to be conducted. Anyhow, it is important that all three aspects are included into a model of eye movements and visual attention allocation.

It turns out, that using more abstract stimuli in a search experiment instead of realistic scenes is helpful to glean insights which would otherwise be hard, if not impossible, to prove. However, it is apparent that the stimuli must not be too abstract. It is important that the stimuli still convey the general picture pertaining to the task. This has been achieved in the conducted experiment by using actual targets and a setup, which despite its simplicity, was reminiscent of an urban scene. Previous experiments using abstract stimuli did not

have a chance to capture the influences of semantically relevant scene locations. In these cases, the stimuli had been too abstract to be able to contain information of that kind.

It is important that the stimuli contain meaningful task-related information in order to allow task-dependent influences to be captured. It is not sufficient to engage the processing of task dependence by just defining target features if one wants to examine top-down influences on search performance, search behavior, and eye movements. The experiment results clearly show that there are much more task-dependent influences than just target presence. Probably, there are even more factors with respect to the task than examined here. Possibly, there are also other top-down aspects not covered by the presented experiment.

The comparison of the relevance map with the eye fixations collected in the second experiment shows that the predictive power of the relevance is very good. The average AUC of 72% exceeds the values reported for the predictive power of the Itti saliency map as well as an improved model based on this saliency map (Einhäuser, Spain, & Perona, 2008). This indicates the importance of considering the semantic relevance of scene locations for eye fixation prediction. Although the relevance maps do very well in predicting eye fixations, it is also very clear that the bottom-up aspects cannot be neglected. A combination of salience maps with relevance maps increases the average AUC even more (74%). This observation is in accordance with the findings of the first experiment which shows that a visual salient distractor can draw the eyes at any time. This finding, however, also indicates that the effects do not just simply add up. This suggests that more sophisticated combination strategies could be employed to further increase the quality of this combined map.

Very interestingly, the two different implementations of salience maps, the original implementation of L. Itti et al. (1998), and the re-implemented version which includes a few changes to the model differ considerably in prediction performance. Most likely, this is due to the normalization scheme which is part of the original visual attention model. This normalization scheme promotes only the strongest peaks of the saliency map and the

intermediate maps feeding into the salience map. The results of the first experiment, however, show that intermediate levels of saliency have a stronger influence on eye movements than very high saliency locations do. These intermediate saliency levels are suppressed by the normalization scheme of the original model, thus leading to inferior prediction performance.

In order to compute the relevance map, a three-dimensional simulation environment with three-dimensional graphical output is required. However, three-dimensional high fidelity graphical output showing the state of a simulated battle is not ubiquitous in military simulations that are used for analysis purposes, yet. One rather recently developed military simulation, IWARS, already provides a similar feature. Also the simulation environment 'Real World', which is designed to be used as analysis tool and as training device, provides three-dimensional graphical output. This seems to show a current trend specifically for high fidelity simulations focusing on representing individual soldiers. Therefore, it can be expected that more simulation of this kind will emerge in the near future. Training simulations have had this kind of graphical output for a long time since it is integral part of the user interface for the trainees. In recent years, especially game-based military simulations have become popular for training individual soldiers in ground combat. These simulations provide three-dimensional graphical output inherently. Using this output, the presented model could be employed for enhancing the realism of non-player characters to enhancing a trainee's sense of presence and thus could possibly improve the training effectiveness of game-based training systems. Likewise, the model could be used for controlling eye movements, and in later stages, head movements of these non-player characters. Realistically animated avatars will enhance the sense of presence for trainees and can furthermore provide realistic cues about what other people pay attention to and where they are fixating on. The gaze direction of other people tends to influence humans.

The newly created model which predicts eye fixations based on a salience map and a relevance map can now be used to improve target detection mechanisms. Instead of applying the currently available mechanisms to very artificially derived fields of view, they can

now be applied to the eye fixation locations predicted by the combined relevance/salience map. The same locations can be used to create false positives. This will tremendously improve the existing target detection mechanisms by endowing them with more human-like behavior, which they actually are supposed to represent.

These improved target detection mechanisms can be added to constructive simulations as well as training simulations. Although the generation of the relevance maps are computational expensive, they can be computed offline for a given scenario. One relevance map would need to be computed and stored for each waypoint in the simulation scenario. This would be quite a large number of maps, but considering the sizes of current consumer storage devices this does not present a challenge at all.

B. FUTURE WORK

The first experiment of this work shows that a distractor has the most distracting effect at a medium saliency level. The strength of this effect differs between the analyzed metrics, and the results currently do not allow for the determination of the exact level of salience that yields the largest distracting effect. It would be very interesting to determine which level of salience has the largest distracting effect. This can be examined by employing stimuli similar to the ones used in the target and distractor condition of the first experiment of this work. However, it would be necessary to use more levels of distractor salience to get a more detailed understanding of the salience effects on the various metrics. This would also require the use of a more objective measure of saliency.

Another interesting aspect to examine is the difference in eye movement behavior between experts and non-experts. For the search task of this work, a person would be considered an expert if he or she has combat experience in the employed setting, in this case urban combat. Non-experts would be people who do not have combat experience and who also never received special training in target detection or in shoot/no-shoot exercises. The definition of expert and non-expert by Wainwright (2008) seems to be a good starting point for making this distinction. Seeing whether the semantically relevant scene locations

have a different influence on the eye movement behavior of experts or non-experts would be the most interesting outcome of this comparison. The comparison should be performed for the analysis of the first experiment of this work and also for the predictive capabilities of the relevance and the salience maps.

With respect to the modeling aspect of this work, several aspects of the model could potentially be refined. The first aspect pertains to the combination of the salience and relevance maps. In the current model, the final prediction map, which is based on the salience map and the relevance map, is created by additively combining these two maps. So far, no other combination strategies have been explored. This would be an interesting future task. One example of combining the two maps in a different way would be to pick the maximum of each map for every pixel. This approach would represent a competition between the salience map and the relevance map.

Furthermore, it would be very interesting to explore additional inputs for the creation of the relevance map. At the moment, the relevance map is based on the fraction of visible target pixels and on the contrast of the target to the background. For the contrast input, the size of the target is currently neglected. However, it is not hard to conceive that blending in with the environment is not just a function of contrast, but is also modulated by target size. For example, it would be interesting to explore how a relevance map including the influence 'contrast $\times$ target size' might be constructed, and how the prediction performance of such a map would compare to the currently used maps.

Another way of possibly improving the prediction performance of the model would be to apply a non-linear transformation to the salience map. Since the first experiment of this work shows that visual feature based attentional capture is maximal at medium saliency levels, a non-linear transformation which takes this knowledge into account should improve the predictive power of the salience map and therefore the predictive power of the combined salience/relevance map.

One aspect, which is known to influence attention allocation but has not been included into the model presented in this work, is the presence of visual features a scene lo-

cation shares with the target (Pomplun, 2006). This means that locations at which features are present that are similar to target features receive a substantial amount of eye fixations. Including this component into the eye movement model presented in this work should improve its prediction performance. It is not immediately apparent which features to use; however, a starting point would be the features used by Pomplun (2006) and another option would be to explore HMAX-like features as defined by Riesenhuber and Poggio (1999).

This last idea goes hand in hand with the determination of false targets. Although it is not clear when and how false targets are generated, it is obvious that false targets usually share features with the target and are therefore mistakenly perceived as targets. When and how these features are used to determine target presence, either correctly or incorrectly, is a subject for future research. Possibly, one could exploit the Recognition-by-Components theory of Biederman (1987) and decompose the target into its basic shapes. Then, the features of these basic shapes could be learned by a HMAX-like model. This model could be used to then determine if a target is detected or not. Similarly, other object recognition mechanisms could be used to detect targets or individual target elements. The difficulty with this approach is that the goal is different from the usual goal of object recognition mechanisms. Usually, one tries to maximize their detection rate. In this case, however, the object recognition mechanisms have to be tweaked such that their performance closely resembles human performance with respect to hits, misses, and false positives.

Another aspect not yet modeled explicitly is the sequence of eye fixations. So far, the model assigns values to scene locations indicating how strong a location attracts the gaze. This map can be used to generate scan paths. Based on earlier findings that one stimulus image elicits very different scan paths for different observers and even for the same observer over different sessions(Mannan et al., 1997), the generation of fixation sequences must not be deterministic. This can be achieved by interpreting the values at the pixel locations of the map as being indicative of the likelihood to fixate these locations. Based on these likelihoods, a fixation is determined. The likelihood at the fixated location decreases exponentially over time while this location is fixated. Thus, the original prediction map is

modified. In addition, based on the observation that human fixations are of limited lengths (Henderson et al., 1999), a cost is associated with the length of the next saccade. This cost is subtracted from each scene pixel based on its distance from the current fixation location. At any point in time, the next fixation location is determined based on the current probabilities associated with each scene pixel.

The work presented here examined the influence of relevant scene locations on attention allocation only for an urban environment. It is dangerous to just assume that the results can be extrapolated to other environments. Instead, additional experiments with target search in other environments need to be conducted. This will allow the comparison of the search behavior and the role of semantically relevant scene locations across different environments. Possibly, the general influence of hiding locations and locations at which targets blend in well with the background are the same for several different environments. However, it could very well be that for different environments different scene elements have a meaning for the search task. Additionally, the performance of experts and non-experts might differ considerably. The author suspects that the search behavior in urban environments does not differ as much between experts and non-experts as the search behavior for natural environments (e.g., forests). The reason being that most people are familiar with urban environments and not so much with forests and wooded areas. Also, it is common for children to play hide and seek while growing up. This provides everybody with some experience as to how and where to look for somebody hiding. This experience, however, is limited based on the environments in which a child grows up. Since most people live in urban environments, it is not unreasonable to assume that more people are familiar with urban environments, than with natural environments. Experts trained for combat in several different environments, will therefore differ more from non-experts in environments an average person is not very familiar with.

This reasoning also indicates that there are possibly additional individual factors that need to be considered aside from being an expert or not. Examining how strong these

differences are and how, if at all, they can be included into a model of human visual perception is another task left for future work.

One very important factor known to easily capture visual attention is movement. This work, however, exclusively examined static scenes with a static observer at a static viewpoint. The reason for this is that including movements into a model of human visual perception is a very complicated endeavor. Therefore, including movement should be performed in several steps. At first, the observer should remain static and only the targets and distractors should move. The extent of these movements should be gradually increased. To begin with, targets and distractors move slightly up and down or left and right while essentially retaining their position. Then, the targets and distractors move realistically in the environment. The next step in the process of including movement would be to have static targets again, but the viewpoint of the observer moves, with the observer still being static. This movement of the viewpoint would resemble the movement of a player in a first-person-shooter game, but with static targets. After that, the moving viewpoint and the moving targets should be combined. The final step would be to not only have the viewpoint move in the simulation, but also require the observer to physically move in order to make the viewpoint move in the virtual environment. This movement of the observer's viewpoint and moving targets would start to extend the research from pure perceptual modeling towards modeling of action and perception and would also require the modeling of the feedback loop between them.

LIST OF REFERENCES

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*, 62-70.

Bacon, W. F., & Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception & Psychophysiscs*, *55*, 485-496.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 115-147.

Biedermann, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, 143-177.

Born, S., & Kerzel, D. (2008). Influence of target and distractor contrast on the remote distractor effect. *Vision Research*, *48*, 2805-2816.

Brockmole, J. R., Castelhano, M. S., & Henderson, J. M. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 699-706.

Brockmole, J. R., & Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition*, *13*, 99-108.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 753-763.

Cavanagh, P. (2004). Attention routines and the architecture of selection. In M. I. Posner (Ed.), *Cognitive neuroscience of attention* (p. 13-28). New York: The Guilford Press.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*, 28-71.

Darken, C., & Jones, B. (2007). Computer graphics-based target detection for synthetic soldiers. In *Proceedings of brims.*

Darken, C. J. (2007). Computer graphics-based models of target detection: Algorithms, comparison to human performance, and failure modes. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, *4*.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Reviews of Neuroscience*, *18*, 193-222.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*, 1-19.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*, 1-26.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861-874.

Findlay, J. M., & Gilchrist, I. D. (2001). Visual attention: The active vision perspective. In M. Jenkins & L. Harris (Eds.), *Vision and attention.* Springer Verlag.

Frintrop, S. (2006). *Vocus: A visual attention system for object detection and goal-directed search* (J. G. Carbonell & J. Siekmann, Eds.). Springer.

Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., & Anderson, C. H. (1994). Overcomplete steerable pyramid filters and rotation invariance. In *Proceedings IEEE computer vision and pattern recognition* (p. 222-228). Seattle, Washington.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29-36.

Harrington, L. (2009). *Adding urban search to traditional search within combat simulations* (Tech. Rep.). U.S. Army Material Systems Analysis Activity.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*, 498-504.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: Insights into mind and brain* (1st ed.). Amsterdam: Elsevier.

Henderson, J. M., McClure, K. K., Pierce, S., & Schrock, G. (1997). Object identification without foveal vision: Evidence from an artificial scotoma paradigm. *Perception & Psychophysics*, *59*(3), 323-346.

Henderson, J. M., Weeks, Jr., P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 210-228.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysiscs*, *57*(6), 787-795.

Hollingworth, A. (2006). Visual memory for natural scenes: Evidence from change detection and visual search. *Visual Cognition*, *14*, 781-807.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 113-136.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature*, *394*, 575-577.

Itti. (2003). Visual attention. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (p. 1196-1201). MIT Press.

Itti, L., Dhavale, N., & Pighin, F. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. In B. Bosacchi, D. B. Fogel, & J. C. Bezdek (Eds.), *Proc. spie 48th annual international symposium on optical science and technology* (Vol. 5200, p. 64-78). Bellingham, WA: SPIE Press.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489-1506.

Itti, L., & Koch, C. (2001a). Computational modeling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194-203.

Itti, L., & Koch, C. (2001b). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, *10*, 161-169.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, *20*(11), 1254-1259.

Kandel, E. R., Schwartz, J. H., & Jessel, T. M. (2000). *Principles of neural science* (Fourth Edition ed.). McGraw-Hill.

Klein, R., Kingstone, A., & Pontefract, A. (1992). Orienting of visual attention. In K. Rayner (Ed.), *Eye movements and visual cognition* (p. 46-65). New York: Springer-Verlag.

Klein, R. M. (2004). On the control of visual orienting. In M. I. Posner (Ed.), *Cognitive neuroscience of attention* (p. 29-44). The Guilford Press.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219-227.

Kunar, M. A., Flusberg, S., Horowitz, T. S., & Wolfe, J. M. (2007). Does contextual cuing guide the deployment of attention? *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 816-828.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception and Psychophysiscs*, *2*, 547-552.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation patterns made during brief examinaton of two-dimensional images. *Perception*, *26*, 1059-1072.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*(2), 205-231.

Navalpakkam, V., & Itti, L. (2006). Top-down attention selection is fine grained. *Journal of Vision*, *6*, 1180-1193.

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, *171*, 308-311.

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. In *Proceedings of the IEEE international conference on image processing* (Vol. I, p. 252-256). Barcelona, Spain.

Peters, R. J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proc. IEEE conference on computer vision and pattern recognition.*

Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research*, *46*, 1886-1900.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-422.

Rayner, K., & Pollatsek, A. (1992). Eye movements and scene perception. *Canadian Journal of Psychology*, *46*, 342-376.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019-1025.

Saida, S., & Ikeda, M. (1979). Useful field size for pattern perception. *Perception and Psychophysiscs*, *25*, 119-125.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, *5*, 195-200.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proceedings of the international conference on image processing.*

Tatler, B. W., Baddeley, R. J., & Glichrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*, 643-659.

Theeuwes, J. (2004). Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review*, *11*(1), 65-70.

Theeuwes, J., Kramer, A. F., Hahn, S., & Irwin, D. E. (1998). Our eyes do not always go where we want them to go: Capture of the eyes by new objects. *Psychological Science*, *9*, 379-385.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*, 766-786.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97-136.

Wainwright, R. K. (2008). *Look again: an investigation of false positive detections in combat models*. Unpublished master's thesis, Naval Postgraduate School.

Wolfe, J. M. (1994). Guided search 2.0, a revised model of visual search. *Psychonomic Bulletin and Review*, *1*, 202-238.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science*, *9*, 33-39.

Yarbus, A. L. (1967). *Eye movements and vision* (L. A. Riggs, Ed.). New York: Plenum Press.

Zhaoping, L., & Dayan, P. (2006). Pre-attentive visual selection. *Neural Networks*, *19*, 1437-1439.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia
2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California
3. Dr. Christian Darken
   Naval Postgraduate School
   Monterey, California
4. Dr. Tony Ciavarelli
   Naval Postgraduate School
   Monterey, California
5. Dr. Rudolph Darken
   Naval Postgraduate School
   Monterey, California
6. John Hiles
   Naval Postgraduate School
   Monterey, California
7. Dr. Thomas Lucas
   Naval Postgraduate School
   Monterey, California
8. Dr. Kevin Squire
   Naval Postgraduate School
   Monterey, California
9. Dr. Timothy Chung
   Naval Postgraduate School
   Monterey, California
10. Dr. Nita Miller
    Naval Postgraduate School
    Monterey, California
11. Dr. Lawrence Shattuck
    Naval Postgraduate School
    Monterey, California

12. LTC Dr. Dave Hudak
    TRADOC Analysis Center
    Monterey, California

13. Jack Jackson
    TRADOC Analysis Center
    Monterey, California

14. MAJ Dr. Paul Evangelista
    TRADOC Analysis Center
    Monterey, California

15. Patrick Jungkunz
    Naval Postgraduate School
    Monterey, California